N° d'enregistrement
au C.N.R.S.

THESE DE DOCTORAT D'ETAT ES SCIENCES APPLIQUEES

présentée

à l'UNIVERSITE PARIS VI

par

Jean-Sylvain LIENARD

pour obtenir

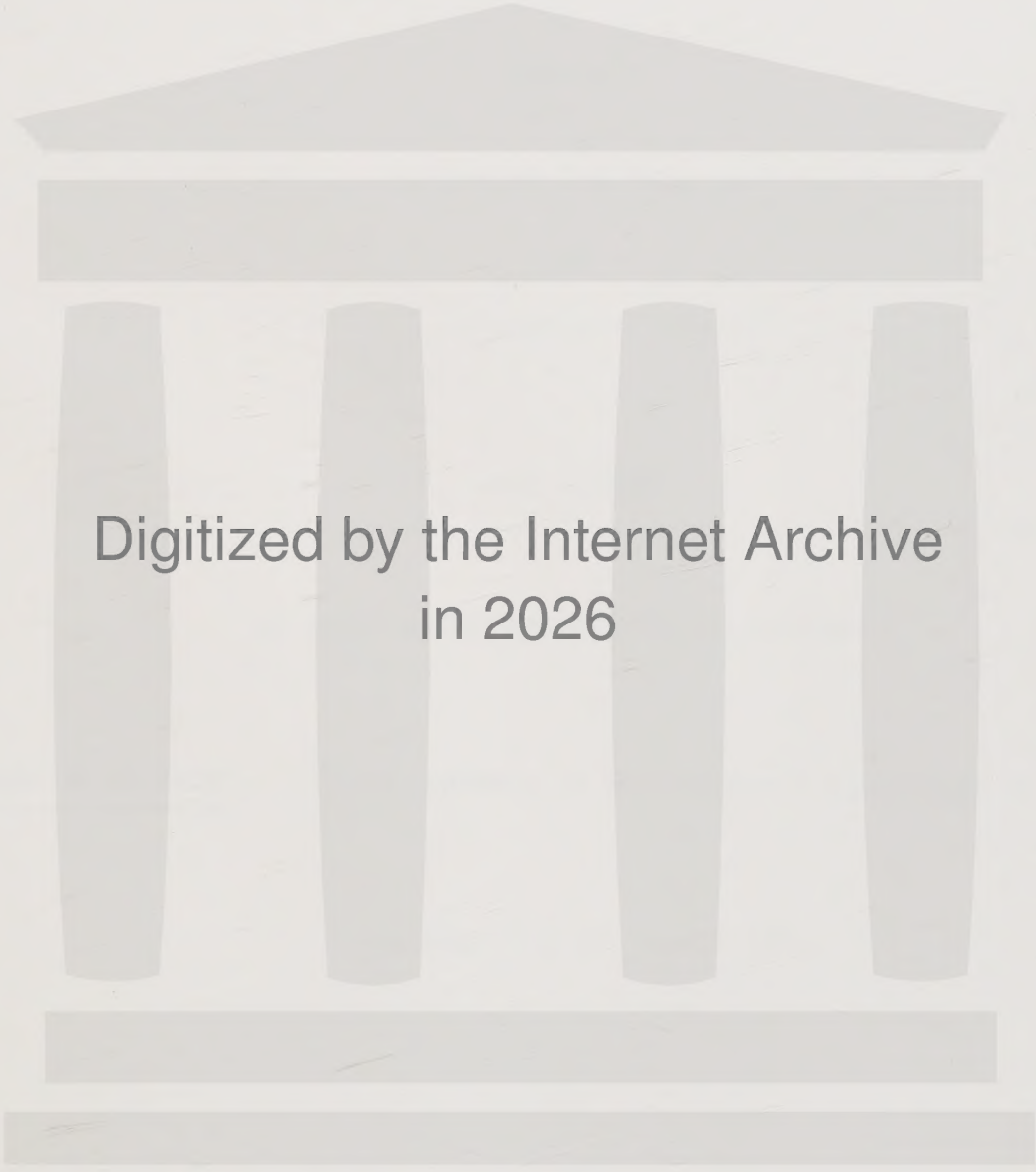
le grade de DOCTEUR ES SCIENCES APPLIQUEES

SUJET DE LA THESE : Analyse, synthèse et reconnaissance automatique de la parole.

Soutenue : le 21 Avril 1972

devant la Commission d'Examen :

MM. L. GAUTHIER Président
L. MALAVARD Examinateurs
R. SIESTRUNCK
J.C.SIMON
E. LEIPP


Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41534912_0004

N° d'enregistrement
au C.N.R.S.

THESE DE DOCTORAT D'ETAT ES SCIENCES APPLIQUEES

présentée

à l'UNIVERSITE PARIS VI

par

Jean-Sylvain LIENARD

pour obtenir

le grade de DOCTEUR ES SCIENCES APPLIQUEES

SUJET DE LA THESE : Analyse, synthèse et reconnaissance automatique de la parole.

Soutenue : le 21 Avril 1972

devant la Commission d'Examen :

MM. L. GAUTHIER Président
L. MALAVARD Examinateurs
R. SIESTRUNCK
J.C.SIMON
E. LEIPP

UNIV.
LEIDEN
BIBL.

AVANT - PROPOS

Depuis huit ans, la doctrine de la parole forgée au Laboratoire d'Acoustique de l'Université PARIS VI (département de Mécanique) n'a cessé de se développer : actuellement un prototype industriel de synthétiseur autonome est en chantier, ainsi qu'un analyseur expérimental destiné à la reconnaissance automatique. Il me semble nécessaire de retracer ici très brièvement les grandes lignes de cette évolution, et de préciser la place que j'y ai occupée.

C'est donc en 1964 que le Laboratoire d'Acoustique commença à s'intéresser aux problèmes spécifiques de la parole. Nous avions constitué, M. LEIPP, Mlle CASTELLENGO et moi-même, un petit groupe d'étude qui s'entraînait à déchiffrer les sonagrammes de parole, en mettant à profit l'expérience acquise en spectrographie musicale.

M. LEIPP posa les bases d'une doctrine originale de la parole : les schémas que nous mettions en évidence à partir des sonagrammes représentaient l'information sémantique de la parole, et relevaient de la théorie de la Forme. Le phonème n'avait pas d'existence physique. Mais la Forme sémantique du message pouvait être considérée comme la juxtaposition de Formes élémentaires répérées par deux symboles phonétiques.

Chacun de nous entreprit, en vue d'une comparaison ultérieure, la constitution d'un "dictionnaire" de ces digrammes phonétiques, ainsi qu'une étude de leur répartition dans des textes variés.

Un premier synthétiseur, l'Icophone I, construit en 1965 par le Laboratoire d'Electronique du département de Mécanique, permit de vérifier la validité de ces idées. Il fut suivi en 1967 de l'Icophone II à lecture optique par transparence, qui permettait une synthèse véritable de la parole; celle-ci était dessinée à la main, en juxtaposant les formes élémentaires requises, extraites du dictionnaire. Les échantillons de cette synthèse sans analyse préalable présentés à Grenoble en avril 1967 ont précédé de plus d'un an les premières réalisations connues de "synthèse par règles" d'intelligibilité comparable.

La méthode montra ses qualités intrinsèques en 1968, lorsque l'Icophone III fut connecté à l'ordinateur du Centre de Calcul Analogique du CNRS : extrême simplicité, commande phonétique (50 bits/seconde), intelligibilité totale, vocabulaire illimité, réponse immédiate. Cet appareillage donna lieu à un brevet international déposé par l'Agence Nationale de Valorisation de la Recherche, et intéressa les responsables de la Direction des Recherches et Moyens d'Essais, qui nous encouragèrent par deux contrats successifs axés à long terme sur la reconnaissance automatique de la parole.

Tout en approfondissant l'étude spectrographique de la parole et l'utilisation de l'Icophone, j'ai eu durant cette période l'occasion de m'intéresser à l'analyse-synthèse de Fourier d'un son permanent, à la statistique phonétique, aux résonateurs, à la synthèse purement mécanique de la parole (reconstitution de la machine parlante de KEMPELEN), et à d'autres sujets de l'acoustique. Depuis 1969 j'ai pu travailler sur les Icophones à commande numérique, et amorcer certaines expérimentations : comparaison d'un élément phonétique à tous les autres, tracé de la courbe de stabilité d'un message en parole synthétique, anamorphose fréquentielle de la voix synthétique, création automatique de mots aléatoires, etc.

...../

J'ai pu alors constater que la courbe de stabilité contenait toute l'information utile à la segmentation et la normalisation des éléments phonétiques, et constituait donc un atout essentiel pour la reconnaissance. L'intérêt du Comité de Recherche en Informatique pour ces idées nouvelles s'est traduit par un contrat portant sur la simulation numérique du processus de reconnaissance, dont la présente thèse expose les premiers résultats.

L'introduction en machine d'un corpus de parole réelle, ainsi que la nécessité de pratiquer sous forme numérique une analyse comparable à celle du Sonagraphe, furent l'objet de nombreuses difficultés. Mais, ce pas franchi avec, en Informatique, l'aide efficace de M. MLOUKA, la segmentation, l'apprentissage et la reconnaissance des éléments phonétiques prirent un tour extrêmement encourageant. L'important corpus de parole disponible en machine m'a permis d'étudier diverses anamorphoses, en vue de réaliser l'adaptation au locuteur. J'ai également effectué des essais de reconnaissance globale par programmation dynamique, avant même d'apprendre que certains chercheurs japonais et soviétiques travaillaient dans la même voie.

L'originalité des idées, des méthodes et du vocabulaire adoptés par notre groupe depuis le début des recherches sur la parole, m'a conduit à tenter un exposé assez vaste, sinon exhaustif. La dernière partie, dans laquelle ma contribution personnelle est plus importante, me semble en effet inséparable de l'ensemble, et j'ai cherché à la situer dans son véritable contexte, celui de la perception des structures acoustiques.

Je souhaite que Monsieur LEIPP, Maître de Recherches au C.N.R.S., dont l'esprit d'invention et le dynamisme sont à l'origine de mon intérêt pour l'Acoustique, puisse voir dans mon travail une expression convenable de ce qu'il ma enseigné.

Je suis très reconnaissant à Monsieur le Professeur SIESTRUNCK, Directeur du Laboratoire d'Acoustique, de la confiance qu'il m'a toujours témoignée, ainsi que de ses conseils et de sa patience infinie.

L'ambiance particulière qui règne dans notre équipe, qui est faite de confiance et d'efficacité, est également due à Monsieur SAPALY, Maître de Conférences à l'Université PARIS VI, responsable de l'étude et de la réalisation des divers appareillages electroniques que nous utilisons.

Que Monsieur le Professeur MALAVARD et Monsieur RENARD, Maître de Recherches au C.N.R.S., qui, en nous ouvrant les portes du Centre de Calcul Analogique, ont permis le développement numérique de l'Icophone, trouvent ici un gage de notre collaboration passée et future.

Monsieur le Doyen GAUTHIER, qui a toujours encouragé et suivi les travaux du Laboratoire d'Acoustique, m'a fait un grand honneur en acceptant la présidence du jury.

Je dois à Monsieur le Professeur MOLES l'initiation aux notions fondamentales de la théorie de l'Information et de la théorie de la Forme, ainsi qu'une certaine vision du monde des idées.

...../

Je remercie Messieurs M.Y. BERNARD, A. DIDIER, A. FOURNIER et J. SALMON, Professeurs au Conservatoire National des Arts et Métiers, qui ont compris l'attrait que présentait pour moi la recherche expérimentale en Acoustique, et ont fait en sorte que je puisse y prendre une part active.

Mademoiselle CASTELLENGO, Assistante au Laboratoire d'Acoustique, s'est montrée la plus habile de notre groupe dans la confection du dictionnaire de synthèse; son aide et son amitié ne se sont jamais démenties. Monsieur TEIL, Ingénieur au C.N.R.S., m'a de longue date facilité l'accès des ordinateurs, et prêté main-forte pour la programmation. Monsieur QUINIO, actuellement Ingénieur à la Société T.I.T.N., a consacré beaucoup de temps et d'efforts à l'adaptation numérique de l'Icophone. Monsieur MLOUKA, expert en traitement de l'information, a mis son enthousiasme et son dévouement dans les programmes de reconnaissance automatique. J'espère que tous verront dans ma thèse un résultat du travail commun.

De nombreuses autres personnes m'ont fourni idées, aide ou encouragement. Je pense notamment à Messieurs les Professeurs P. ALAIS et J.C. SIMON; Madame BOREL-MAISONNY; Monsieur MOUTET, Monsieur DESMARAIS; Madame CALLENS et Madame LALIMAN; Messieurs COTIN et DUBOC; Messieurs CALINET et BRAURE. Et je n'oublie pas ici les personnes de mon proche entourage, dont la compréhension m'a permis de mener à bien ce travail.

PARIS, le 16 Mars 1972

I - INTRODUCTION

I.1 GENERALITES

Le domaine de l'acoustique est à la fois trop vaste et trop cloisonné pour permettre une délimitation exacte de notre champ d'investigation. Trop vaste, car nous ne traiterons ni d'acoustique physique ou mathématique, ni d'acoustique sous-marine. Trop cloisonné, car les problèmes relatifs à la parole ont des origines ou implications variées : dans la psycho-physiologie de l'audition et de la phonation, dans la mécanique vibratoire, dans la linguistique, l'électronique, l'informatique, et la science des communications. Nous devrons donc emprunter à ces sciences ou techniques connexes des concepts, des méthodes et des outils, emprunts nécessairement maladroits ou incomplets aux yeux des spécialistes que nous prions ici de nous excuser.

Malgré une large utilisation des moyens offerts par l'Informatique, cette étude relève essentiellement de l'Acoustique Musicale, telle qu'elle est pratiquée au Laboratoire d'Acoustique de l'Université de PARIS VI. Les idées directrices, les moyens de travail, les motivations, sont ceux-là mêmes qui apparaissent dans les études relatives aux instruments de musique ou à la perception du bruit (cf. bib. I-1, I-2, et Bulletins du Groupe d'Acoustique Musicale). Il s'agit d'une acoustique perceptive, fortement pragmatique, dont on peut comme suit résumer la doctrine : pour mieux communiquer avec ses semblables, pour mieux utiliser son environnement, ou pour sa simple délectation personnelle, l'homme utilise depuis des millénaires divers signaux acoustiques, bruit, parole, musique, qu'il a façonnés pour leur donner une efficacité maximum, eu égard à la perception. L'étude de ces signaux et de leur production est une voie détournée, mais sûre, menant à la compréhension de certains mécanismes perceptifs. Cette connaissance permet en retour d'aborder de nombreux problèmes pratiques, allant de la lutherie à la rééducation orthophonique, l'insonorisation et la composition musicale, sans oublier naturellement la synthèse et la reconnaissance automatique de la parole.

La parole est un cas particulier du message acoustique; mais tous les problèmes de l'acoustique musicale y sont présents : mécanique, psycho-physiologie, communication etc... Par ailleurs, le message parlé a pris une importance croissante dans notre société depuis le siècle dernier, avec l'apparition du téléphone, de la radiodiffusion et des moyens d'enregistrement. Cette considération devient d'une actualité brûlante si l'on envisage conjointement l'évolution du machinisme en général, et de l'informatique en particulier. En effet, l'intérêt essentiel du message parlé, par rapport au message écrit, est sa spontanéité : la parole est le plus rapide et le plus complet des moyens de communication humains. Or nombre d'applications de l'informatique requièrent un véritable dialogue entre l'homme et ses machines, qui jusqu'à présent s'effectue par l'intermédiaire du clavier alphanumérique dans un sens et d'imprimantes ou écrans cathodiques dans l'autre. La commande vocale rendrait plus rapide et plus souple l'entrée en machine; la réponse vocale permettrait, entre autres avantages, de libérer l'opérateur actuellement rivé à sa console de visualisation, et d'oublier à mesure une grande quantité d'information inutile.

Si de tels appareillages étaient réalisables - et nous verrons plus loin que cette proposition est légitime, bien que l'adéquation des moyens humains soit encore, et pour longtemps, hors de portée - ils trouveraient immédiatement d'autres applications économiques ou culturelles. La gestion automatique, le tri postal, la conduite de certaines machines, les techniques de télétransmission et d'enseignement audio-visuel, en bref toutes les activités automatisées pouvant utiliser le message vocal y trouveraient leur compte.

...../

Pour en arriver là, il faut encore vaincre de nombreuses difficultés, dont la plupart sont d'ordre fondamental plus que technologique : dans quels aspects du message acoustique faut-il chercher l'information utile, et quels en sont les paramètres ? Comment compenser les innombrables variations individuelles du signal physique, dont le champ de liberté s'étend dans trois dimensions ? Comment reconnaître la parole malgré les inévitables lacunes et défauts de prononciation, etc... L'étude qui suit a l'ambition d'apporter quelques éléments de réponse à ces questions, en s'appuyant sur l'expérimentation acoustique et la simulation numérique. Elle s'inscrit dans une série de travaux menés au Laboratoire d'Acoustique avec un double objectif, d'une part scientifique - connaissance par analyse et synthèse des structures acoustiques de la parole; compréhension fonctionnelle de mécanismes mis en jeu par l'audition et la phonation - et d'autre part technologique : mise au point de systèmes aptes à produire, recevoir, et transmettre de l'information sous forme de parole.

I.2. PLAN DE L'ETUDE

Le message parlé est fait pour être produit et utilisé par l'homme dans son milieu naturel. Il s'est élaboré en fonction de contraintes imposées notamment

- par l'appareil émetteur (organe phonatoire), dont la complication est très grande et les proportions extrêmement variables selon les individus.
- par l'appareil récepteur (oreille) indissociable du cerveau, et dont le fonctionnement n'est pas encore élucidé, du moins pour les signaux complexes évolutifs.
- par les besoins de la communication humaine, qui s'équilibrent selon le principe du moindre effort.
- par l'environnement et le mode de vie des individus, qui impliquent certaines conditions acoustiques telles que la forme et l'intensité du bruit de fond, la portée de la voix etc...

L'évolution des deux derniers points, qui dépendent du niveau et du type de civilisation, se traduit par une évolution permanente de la langue et de sa prononciation. Ainsi, dans notre société, se créent des mots nouveaux répondant à l'extension des activités humaines; mais simultanément la langue s'uniformise et se simplifie, comme en témoigne la quasi-disparition de certaines nuances phonétiques, et même s'abâtardit dans certains jargons (celui de l'informatique par exemple).

Les conditions matérielles de la vie ont également leur importance : on ne parle pas de la même manière à la campagne et à la ville, dans un appartement ou dans le métro. L'environnement impose un bruit de fond spécifique, la société impose une distance donnée entre individus en activité : la portée de la voix varie en conséquence, ainsi que le timbre, le rythme, l'articulation, et le choix des messages à transmettre.

Des multiples aspects du message parlé nous retiendrons surtout la plasticité nécessaire pour remplir son rôle dans la communication humaine : le message parlé doit pouvoir être émis et reçu par le plus grand nombre d'individus, tous différents, et résister à des conditions variées de transmission et d'utilisation, ce qui justifie l'énorme redondance de l'information sémantique.

Le parallélisme entre langage parlé et langage écrit n'est qu'illusoire, et nous devrons nous en défier. La notation en mots séparés par des blancs, par exemple, n'a aucune correspondance acoustique : le message parlé est continu entre deux ponctuations.

..../

De même l'écriture des langues occidentales, qui utilise des symboles alphabétiques discrets, n'a qu'une relation très lâche avec la réalité acoustique : la notion de phonème est tout à fait abstraite.

L'étude du message parlé peut emprunter diverses voies d'approche. L'analyse spectrale et la théorie de l'Information nous fourniront des notions physiques et mathématiques sur la structure du signal de parole, indispensables pour établir par la suite des corrélations valables entre les mesures objectives et les observations. La parole, faite pour être entendue et perçue, est sans aucun doute déterminée très largement par les caractéristiques de l'audition humaine; nous rappellerons, parmi ces caractéristiques, celles qui nous semblent essentielles. Ceci nous mènera à la théorie de la Forme, qui complète la théorie de l'Information dans le domaine perceptif. Les contraintes dûes à la phonation et à la phonétique, bien qu'importantes, nous paraissent secondaires par rapport à celles qui proviennent de l'audition et de la perception : la preuve en est dans la facilité qu'ont certains animaux (le mainate par exemple) à imiter parfaitement la parole humaine, et dans le fait que l'on sache actuellement beaucoup mieux synthétiser la parole que la reconnaître par des moyens techniques. Nous nous attarderons quelque peu sur certains aspects de la statistique phonétique, avant de faire un rapide tour d'horizon sur les synthétiseurs de parole et leur utilisation scientifique et industrielle.

Les premiers essais de synthèse de la parole datent du 18ème siècle, mais il a fallu attendre l'apparition de l'électronique pour savoir mettre au point de véritables synthétiseurs. Quelques machines parlantes seront décrites dans le chapitre " Synthèse ", à titre surtout de modèles phonatoires. Les principaux types de synthétiseurs modernes y figureront également, avec leurs qualités et leurs défauts éventuels, dans la mesure où ceux-ci ont quelque rapport avec les hypothèses précédentes. Nous décrirons ensuite les différentes versions de l'Icophone et les expérimentations qu'il nous a permis de mener à bien.

La reconnaissance automatique de la parole constituera la dernière partie de ce travail. Après avoir dégagé les principales tendances de la recherche dans ce domaine, en prenant pour exemples quelques réalisations-types, nous définirons notre propre orientation à partir de l'expérience acquise en analyse et en synthèse. Nous décrirons ensuite un processus de reconnaissance entièrement simulé sur ordinateur. Après de nombreux essais, deux options ont été retenues et donnent lieu à des investigations simultanées. L'une concerne la reconnaissance phonétique, phonatome par phonatome, et l'autre la reconnaissance par mots, susceptible de conduire plus rapidement à une réalisation simple. Les deux processus ont en commun les idées développées dans les chapitres précédents : l'anamorphose, par exemple, est à la base de l'adaptabilité à un locuteur quelconque; la notion de phonatome, élément de mouvement des organes phonatoires, conduit à étudier la stabilité temporelle du squelette sémantique dont les extrêma successifs permettent la segmentation et la normalisation temporelle.

Les taux de reconnaissance que nous pourrons annoncer à la fin de ce travail ne devront nullement être considérés comme les performances optimales d'un système de reconnaissance, mais comme une confirmation du bien-fondé des hypothèses. Tout en connaissant les limites des systèmes conçus à partir de ces idées, qui n'incluent pas encore certaines activités cérébrales ou sensorielles fondamentales, nous pouvons espérer que leur approfondissement se révélera fructueux dans un grand nombre de problèmes concernant la parole.

...../

II - ANALYSE SPECTRALE ET INFORMATIONNELLE EN ACOUSTIQUE

II.1 TRANSFORMATION DE FOURIER

Il n'est pas question d'aborder ici l'ensemble des méthodes de traitement des signaux. Nous nous bornerons à rappeler quelques traits fondamentaux de la transformation de FOURIER et des méthodes d'analyse spectrale utilisées par la suite, y compris leur version discrète adaptée au traitement numérique.

II.1.1. Bases mathématiques de l'analyse spectrale

Nous considérerons dans la suite un signal acoustique $f(t)$, fonction uniquement du temps et non de l'espace. Cette fonction, qui peut représenter la variation de la pression acoustique en un point donné, est étudiée dans la pratique par l'intermédiaire de son analogue électrique, qui est généralement la différence de potentiel alternative recueillie aux bornes d'un microphone. C'est dire que nous laisserons de côté :

- la composante continue du signal,
- les phénomènes liés à la directivité et à l'efficacité du microphone,
- l'aspect multi-dimensionnel ou " stéréophonique " du signal acoustique.

II.1.1.1. Série de FOURIER

D'après le théorème de FOURIER, toute fonction périodique $f(t) = f(t + T)$ peut être décomposée en une somme de sinusoïdes harmoniques :

$$f(t) = P_o + \sum_{n=1}^{\infty} P_n \sin(n\omega_o t + \varphi_n)$$

T est la période de $f(t)$, en secondes ; $\omega_o = \frac{2\pi}{T}$ est la pulsation (en rd/s), et $F = \frac{1}{T}$ la fréquence (Hz). P_o est la composante continue. $n = 1, 2, \ldots$, est le numéro de l'harmonique; P_n et φ_n sont l'amplitude (unité arbitraire) et la phase (radians) de l'harmonique n.

L'intérêt de la décomposition de FOURIER est de séparer phase et amplitude. Il est équivalent de représenter une vibration périodique par le tracé de la fonction $f(t)$ pendant une période, ou par le tracé des diagrammes de phase et d'amplitude en fonction des numéros des harmoniques (bib. II-1). Or l'oreille semble n'être sensible, dans de larges limites, qu'à la répartition fréquentielle de l'énergie (bib. II-2). Elle ne discrimine pas, en première approximation, deux

...../

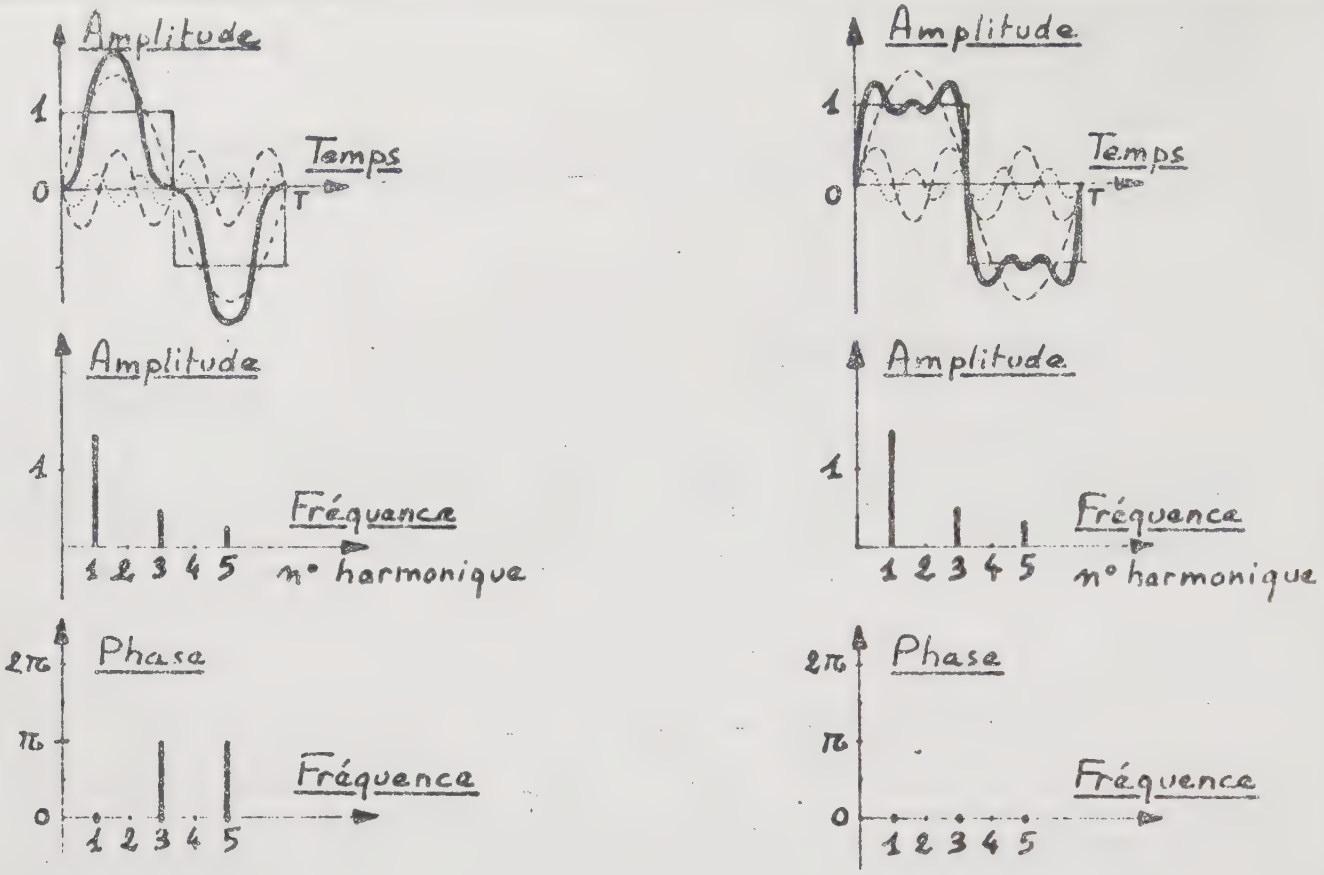


Fig II-1 Représentation temporelle de deux signaux ayant la même composition harmonique, mais des spectres de phase différents . La figure de droite représente un signal carré aproché, reconstitué à partir de ses trois premiers harmoniques.

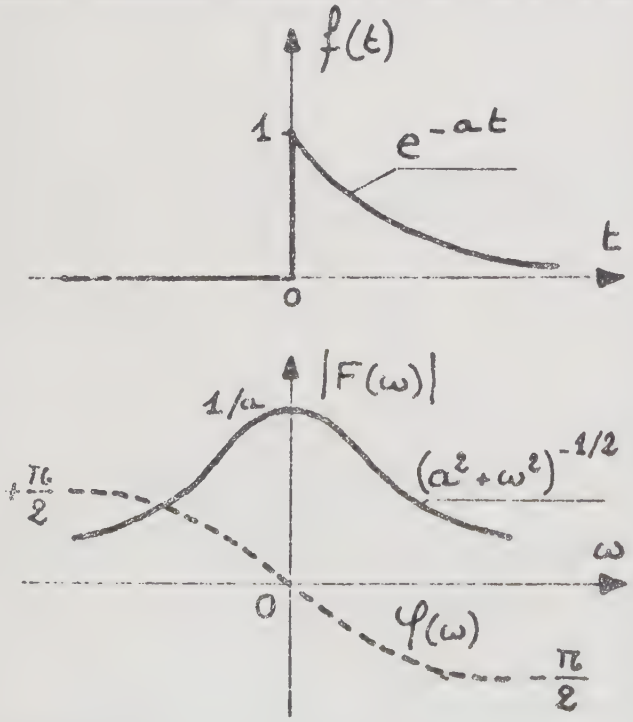


Fig II-2 Transformée de FOURIER d'un signal exponentiel. Le spectre d'amplitude $F(\omega)$ est complexe :

$$F(\omega) = |F(\omega)| . e^{-j\varphi(\omega)}$$

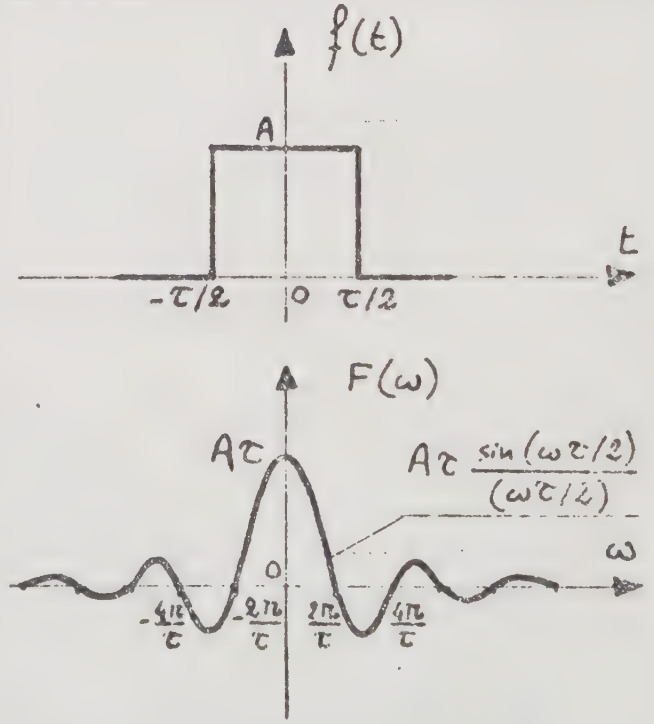


Fig II-3 Transformée de FOURIER d'une impulsion carrée. La fonction est paire; le spectre de phase est identiquement nul, et le spectre d'amplitude est réel.

sons ayant la même composition harmonique et des diagrammes de phase différents, et ayant par conséquent des formes temporelles très différentes (fig. II.1).

Calcul des coefficients :

On peut écrire la série de FOURIER sous une autre forme :

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega_0 t + \sum_{n=1}^{\infty} b_n \sin n\omega_0 t$$

avec

$$a_0 = \frac{1}{T}\int_{t_0}^{t_0+T} f(t)\,dt$$

$$a_n = \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\cos n\omega_0 t\,dt$$

$$b_n = \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\sin n\omega_0 t\,dt$$

Les amplitudes et phases de la première série se calculent à partir de ces coefficients :

$$P_0 = a_0$$

$$P_n = \sqrt{a_n^2 + b_n^2}$$

$$\operatorname{tg}\varphi_n = \frac{b_n}{a_n}$$

Réciproquement :

$$a_n = P_n \cos\varphi_n$$

$$b_n = P_n \sin\varphi_n$$

Notation complexe

Les formules d'Euler et le calcul complexe permettent de simplifier la notation de la série, en condensant les deux séries trigonométriques en une série exponentielle :

$$f(t) = \sum_{n=-\infty}^{+\infty} F_n\, e^{jn\omega_0 t} \qquad -\infty < t < +\infty$$

F_n : coefficient complexe défini par :

$$F_n = \frac{1}{T}\int_{t_0}^{t_0+T} f(t)\, e^{-jn\omega_0 t}\,dt \quad \cdot \quad \begin{cases} n \text{ entier} \\ -\infty < n < +\infty \end{cases}$$

La notation complexe est extrêmement utile et condensée, mais moins proche de la réalité auditive que la première notation. Les correspondances sont les suivantes :

...../

$$F_n = \frac{1}{2}(a_n - j b_n) = \frac{1}{2} P_n e^{-j\varphi_n}$$

$$F_{-n} = \frac{1}{2}(a_n + j b_n) = \frac{1}{2} P_n e^{j\varphi_n}$$

II.1.1.2. Intégrale de FOURIER. (bib. II-3, II-4, II-5)

La série de FOURIER, relative à des signaux périodiques ou quasi-périodiques ne présente pas grand intérêt du point de vue de la communication; en effet, un signal rigoureusement périodique ne transporte aucune information. Par contre l'intégrale de FOURIER permet d'analyser des signaux quelconques en éliminant le paramètre temps au profit du paramètre fréquence. La fonction $f(t)$ possède en général une transformée $F(\omega)$ dans le domaine fréquentiel :

$$F(\omega) = \int_{-\infty}^{+\infty} f(t)\, e^{-j\omega t}\, dt$$

Réciproquement, à une fonction $F(\omega)$ du domaine spectral est associée une fonction $f(t)$ du domaine temporel par la transformée inverse de FOURIER :

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega)\, e^{j\omega t}\, d\omega$$

On note symboliquement :

$$F(\omega) = \mathcal{F}[f(t)] \qquad \text{et} \qquad f(t) = \mathcal{F}^{-1}[F(\omega)]$$

En acoustique la fonction du temps $f(t)$ est réelle; sa représentation $F(\omega)$ dans le domaine spectral est généralement complexe, mais possède de remarquables propriétés de symétrie.

En effet, soient $|F(\omega)|$ le module de la transformée et $\varphi(\omega)$ sa phase par rapport à l'origine : $F(\omega) = |F(\omega)|\, e^{j\varphi(\omega)}$. On montre facilement que F est une fonction paire de ω et φ une fonction impaire. En d'autres termes, il suffit de connaître les valeurs de $F(\omega)$ dans l'intervalle $0 \leqslant \omega \leqslant \infty$ pour savoir reconstruire $f(t)$. Prenons quelques exemples :

Ex. 1 : Transformée d'un signal exponentiel isolé (fig. II-2)

Ce signal est défini par

$$\begin{cases} f(t) = 0 & t < 0 \\ f(t) = e^{-at} & \begin{cases} t \geqslant 0 \\ a > 1 \end{cases} \end{cases}$$

$$F(\omega) = \int_{-\infty}^{+\infty} f(t)\, e^{-j\omega t}\, dt = \int_{-\infty}^{+\infty} e^{-(a+j\omega)t}\, dt = \frac{1}{a+j\omega}$$

...../

D'où $|F(\omega)| = \frac{1}{\sqrt{a^2+\omega^2}}$ et $\varphi(\omega) = -\arctg\left(\frac{\omega}{a}\right)$

Le spectre de densité d'énergie $E(\omega)$ (J.s, Joules par Hertz), est proportionnel au carré du module de la transformée. Pour simplifier les expressions, la constante de proportionnalité est considérée comme égale à 1

$$E(\omega) = |F(\omega)|^2$$

Cette fonction est comme $F(\omega)$ une fonction paire de ω .

Ex. 2 : Transformée d'un signal carré isolé (fig. II-3)

$$\begin{cases} f(t) = 0 & \begin{cases} t < -\frac{\tau}{2} \\ t > \frac{\tau}{2} \end{cases} \\ f(t) = A & -\frac{\tau}{2} \leq t \leq \frac{\tau}{2} \end{cases}$$

$$F(\omega) = \int_{-\tau/2}^{\tau/2} A\, e^{-j\omega t}\, dt = A\tau\, \frac{\sin\frac{\omega\tau}{2}}{\frac{\omega\tau}{2}}$$

Dans ce cas le spectre de phase $\varphi(\omega)$ est nul; ceci provient de la parité de $f(t)$. La densité d'énergie est naturellement toujours positive. On voit que pour un signal de durée $\tau = 1$ ms, par exemple, les maxima d'énergie (pôles de la transformée) se trouvent aux fréquences 0; 1500; 2500; 3500 Hz etc..., et les minima d'énergie (zéros de la transformée) se trouvent à 1000; 2000; 3000 Hz etc...

Sans multiplier les exemples, rappelons que la transformée d'une impulsion de DIRAC $\delta(t)$ est une constante ("bruit blanc") et que la transformée d'un signal périodique est un ensemble de valeurs discrètes, analogue à celui qui résulterait d'une décomposition en série de FOURIER.

Lorsque le signal n'a pas une énergie finie (cas des signaux non limités dans le temps) la notion de densité d'énergie spectrale perd son sens. On utilise alors une fonction comparable, le spectre de densité de puissance moyenne, ou spectre de puissance $S(\omega)$ défini par

$$S(\omega) = \lim_{T\to\infty} \frac{|F_T(\omega)|^2}{T}$$

$F_T(\omega)$ est le module de la transformée de FOURIER du signal considéré dans un intervalle $\left[-\frac{T}{2}, +\frac{T}{2}\right]$ tendant vers l'infini. Au facteur 2π près, $F_T(\omega)$ s'exprime en X.s ou X par Hertz, (X : unité d'amplitude utilisée pour décrire le signal; par exemple Volt ou Pascal). $S(\omega)$ s'exprime en W.s ou Watt par Hertz. Il existe donc toujours une constante de proportionnalité implicite, égale à 1 , de dimension appropriée. Par exemple, si le signal est dé-

...../

crit par une différence de potentiel (X, Volts), S (ω) est la densité spectrale de puissance dissipée dans une résistance de 1 Ohm.

II.1.1.3. Le théorème de convolution

Si f_1 (t) et f_2 (t) sont deux fonctions du temps possédant des transformées de FOURIER F_1 (ω) et F_2 (ω) , la transformée F (ω) du produit f_1 (t) . f_2 (t) est donnée par :

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F_1(u) \cdot F_2(\omega - u)\, du$$

Cette relation, nommée produit de convolution, est notée symboliquement :

$$F(\omega) = \frac{1}{2\pi} \left[F_1(\omega) \otimes F_2(\omega) \right]$$

Réciproquement, dans le domaine temporel, la fonction f (t) correspondant au produit F_1 (ω) . F_2 (ω) est donnée par

$$f(t) = \int_{-\infty}^{+\infty} f_1(\lambda) \cdot f_2(t-\lambda)\, d\lambda$$

relation notée symboliquement f (t) = f_1 (t) $\otimes$ f_2 (t)

Notons que le produit de convolution est commutatif, distributif et associatif.

II-1.1.4. Autocorrélation et spectre de puissance

L'idée d'autocorrélation d'un signal est extrêmement précieuse, comme nous le verrons par la suite. Elle consiste à comparer le signal à lui-même avec un certain délai τ , que l'on fait varier de $-\infty$ à $+\infty$. Pour chaque valeur de τ , on définit l'autocorrélation $\mathcal{R}(\tau)$ comme la valeur moyenne du produit f (t) . f (t - τ) . D'où

$$\mathcal{R}(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{+T/2} f(t) \cdot f(t-\tau)\, dt$$

Fonction d'autocorrélation et spectre de puissance (ou d'énergie) ne sont que deux manières de décrire le même phénomène, c'est-à-dire d'éliminer les phases. En effet, ces deux fonctions se correspondent dans une transformation de FOURIER :

$$S(\omega) = \mathcal{F}\left[\mathcal{R}(\tau)\right]$$

La figure II-4 donne l'allure du spectre de puissance et de la fonction d'autocorrélation de quelques signaux simples.

...../

Signal	Spectre de puissance	Autocorrélation
Sinusoïde		
Constante		
Impulsion de Dirac		
Signal à bande limitée		

Fig II-4 Correspondance entre trois descriptions possibles de quelques signaux-types.

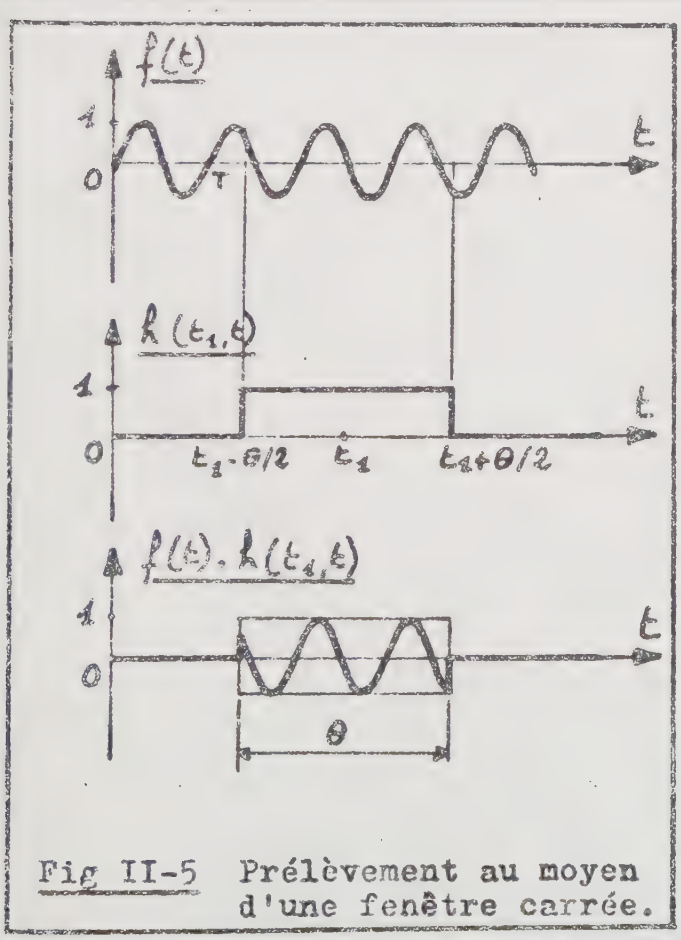


Fig II-5 Prélèvement au moyen d'une fenêtre carrée.

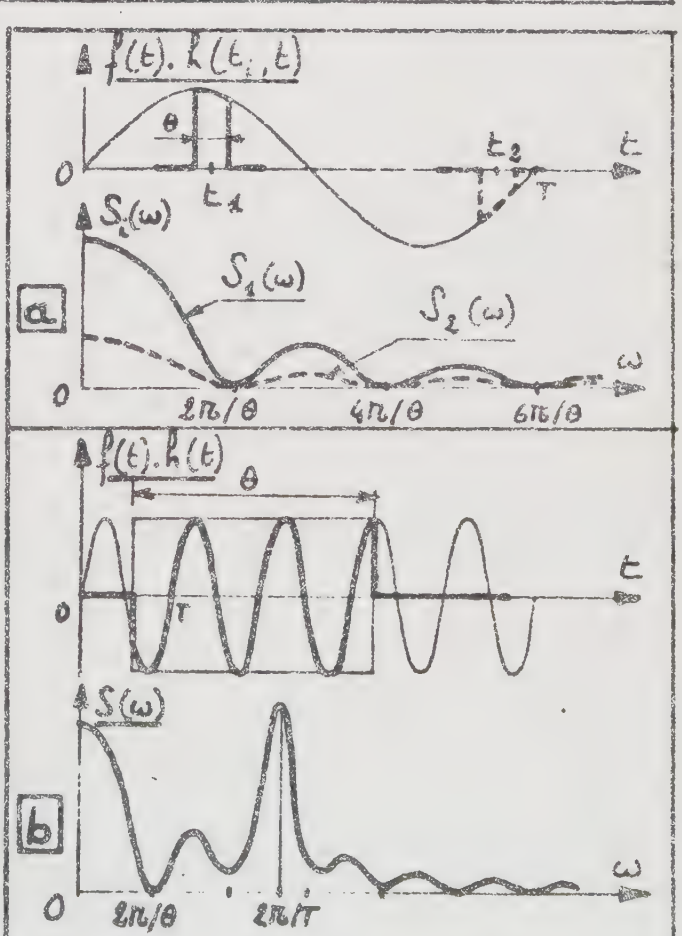


Fig II-6 Influence de la largeur du prélèvement et de sa position sur le spectre du signal. a) Fenêtre très étroite. b) Fenêtre de largeur égale à quelques périodes du signal

II.1.2. Notion de spectre instantané.

II.1.2.1. Définition (bib. II-5, II-6)

Jusqu'à présent nous avons considéré le signal $f(t)$ comme un tout, quelque soit sa durée, et nous lui avons fait correspondre un seul spectre moyen d'amplitude ou de puissance dans le domaine fréquentiel. Mais seuls les changements du signal au cours du temps sont susceptibles de convoyer de l'information. Il faut pouvoir décrire ces changements dans le domaine spectral.

Pour cela nous pouvons considérer une " tranche " du signal, de durée θ. Le prélèvement peut être effectué en multipliant à l'instant t_1, le signal $f(t)$ par une " fenêtre " $h(t_1, t)$ que nous supposerons carrée pour l'instant (fig. II-5) :

$$\begin{cases} h(t_1, t) = 1 & t_1 - \frac{\theta}{2} \leqslant t \leqslant t_1 + \frac{\theta}{2} \\ h(t_1, t) = 0 & \text{partout ailleurs} \end{cases}$$

Le signal résultant possède une transformée de FOURIER que nous appellerons spectre instantané à l'instant t_1. En répétant cette opération à divers instants t_1, on obtient une succession de spectres (d'amplitude ou de puissance) qui reflètent l'évolution temporelle de $f(t)$.

Le spectre instantané est fortement déterminé par la forme et la durée de la fenêtre $h(t_1, t)$. Nous allons détailler quelque peu ce point, qui est au centre du problème d'analyse de la parole.

II.1.2.2. Choix de la fenêtre temporelle (bib. II-5, II-7).

Soient $f(t)$ le signal à analyser, $F(\omega)$ son spectre d'amplitude, $h(t)$ la fenêtre temporelle à l'instant zéro et $H(\omega)$ sa transformée. On exprime souvent $h(t)$ en s^{-1}, de façon que $H(\omega)$ soit sans dimension. Le spectre à court terme à l'instant zéro est obtenu par la convolution

$$\mathcal{F}\left[f(t).h(t)\right] = \frac{1}{2\pi}\left[F(\omega) \otimes H(\omega)\right]$$

Pour étudier l'influence sur l'analyse de la durée θ du prélèvement, supposons que $f(t)$ soit une sinusoïde d'amplitude unité, de période T, de pulsation ω et fréquence F, et $h(t_1, t)$ une fenêtre carrée, de largeur θ (secondes) et de hauteur 1 s^{-1}, appliquée à l'instant t_1 (fig. II-6).

Si $\theta \ll T$ la fonction $\left[f(t) . h(t_1, t)\right]$ se rapproche d'une impulsion dont l'amplitude varie entre $+1$ et -1 selon la valeur de t_1. Le spectre de puissance instantané voit sa valeur maximum passer de 0 à 1 deux fois par période. Notons que la moindre dissymétrie de $f(t)$ ou de $h(t_1, t)$ suffirait à différencier les deux spectres maxima l'un de l'autre. Autrement dit l'analyse des signaux réels avec une fenêtre trop étroite (correspondant à une fenêtre spectrale trop large) ne permet pas de discerner la fréquence dans la forme du spectre; mais la périodicité du signal apparaît dans la sucession des spectres instan-

...../

tanés.

Inversement, si $\theta \gg T$, le spectre instantané se rapproche du spectre moyen du signal, la fréquence apparaît clairement dans la forme du spectre, mais les éventuelles variations temporelles du signal ne sont plus décelées.

Le choix de θ est donc dicté par la bande de fréquence dans laquelle évolue le signal, et par ce que l'on veut observer : un compromis doit être fait entre la résolution temporelle et la résolution fréquentielle de l'analyse.

La forme de la fenêtre temporelle est également très importante. Considérons par exemple la même fonction $f(t) = \sin \omega_0 t$ modulée à l'instant $t_1 = 0$ au moyen de deux fenêtres différentes de même durée θ , l'une $h_1(t)$ rectangulaire, l'autre $h_2(t)$ triangulaire (fig. II-7 a et b). Les deux spectres instantanés $F_1(\omega)$ et $F_2(\omega)$ sont obtenus par convolution :

$$F_1(\omega) = F(\omega) \otimes H_1(\omega)$$

$$F_2(\omega) = F(\omega) \otimes H_2(\omega)$$

Les spectres instantanés sont passablement différents non seulement par leur largeur, mais aussi et surtout par leur forme (amplitude et répartition des rebondissements latéraux), qui est déterminée par la forme de la transformée $H(\omega)$ de la fenêtre $h(t)$.

La figure II-7 c représente l'analyse instantanée effectuée à l'instant $t = 0$ avec une fenêtre particulière dite fenêtre de " HANNING " (de Julius von HANN, météorologiste australien) qui minimise les perturbations spectrales, et que nous utiliserons au chapitre VI :

$$h_a(t) = \frac{1}{2} \left(1 + \cos \frac{2\pi t}{\theta}\right)$$

II.1.2.3. Analyse au moyen de filtres

Au lieu de prélever des "tranches" du signal on peut appliquer celui-ci en continu à l'entrée d'un filtre ou d'un banc de N filtres. Chaque filtre possède une "courbe de réponse" ou représentation dans le domaine spectral $K_n(\omega)$. Il est évident que, lorsque l'opération de filtrage est terminée, le signal filtré est représenté dans le domaine spectral par l'expression

$$F(\omega) = \sum_{n=1}^{N} K_n(\omega)$$

Pour obtenir un spectre évoluant dans le temps nous enregistrons à chaque instant l'amplitude du signal obtenu à la sortie de chaque filtre. L'ensemble de ces signaux $g_n(t)$ donne lieu après détection quadratique à un spectre de puissance à court terme $|F(\omega, t)|^2$. Ce spectre n'est pas le même que celui que nous avons obtenu précédemment à partir d'une fenêtre $h(t)$, car il dépend des caractéristiques des filtres utilisés.

En particulier la réponse impulsionnelle d'un filtre analogique est généralement dissymétrique, différant par là des fenêtres temporelles examinées jusqu'à

..../

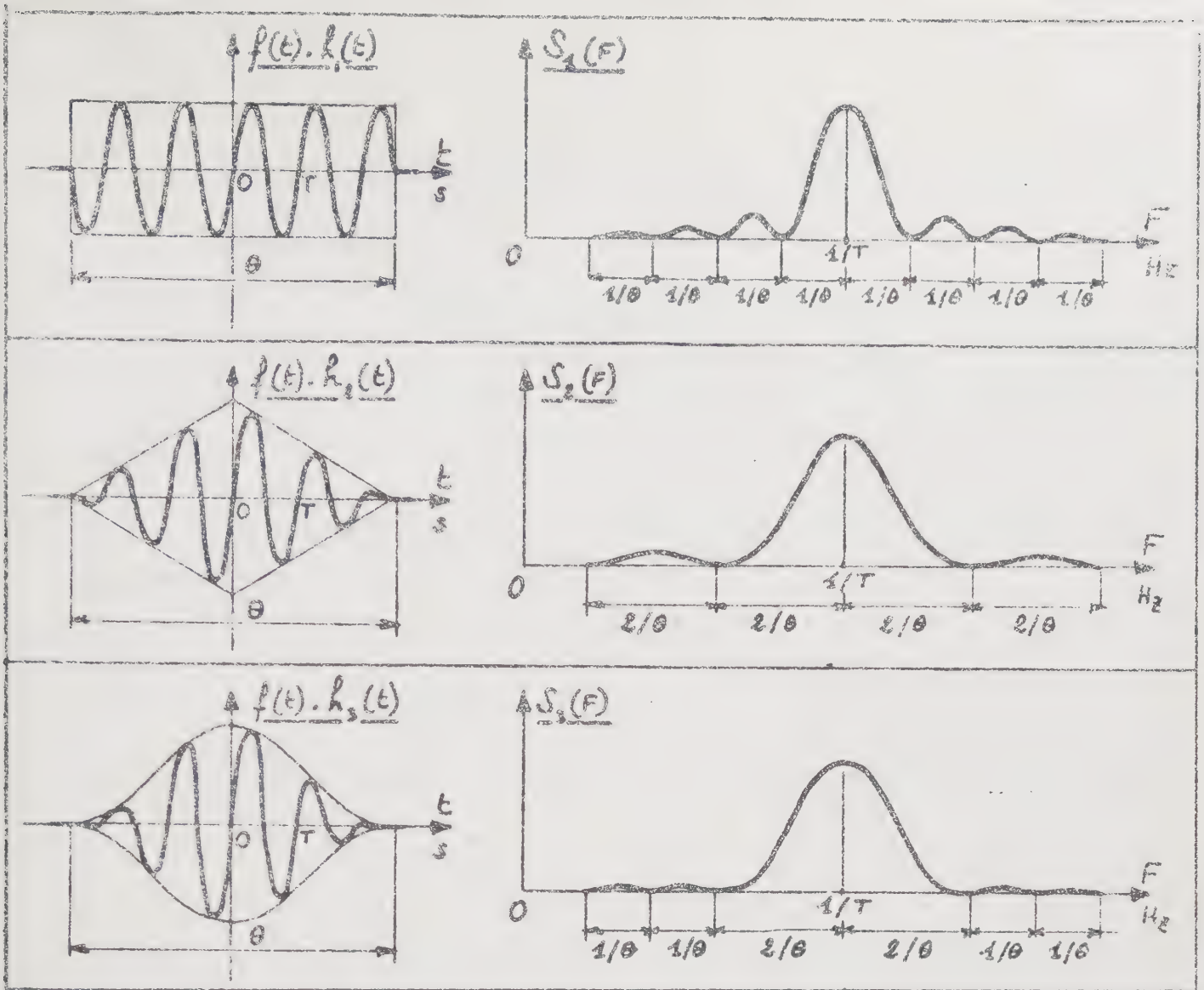


Fig II-7 Action de différentes fenêtres temporelles, de durée θ , sur le spectre instantané d'une sinusoïde de période T. a) Fenêtre carrée. b) Fenêtre triangulaire. c) Fenêtre de HAMMING. Cette dernière réduit considérablement l'amplitude des lobes latéraux.

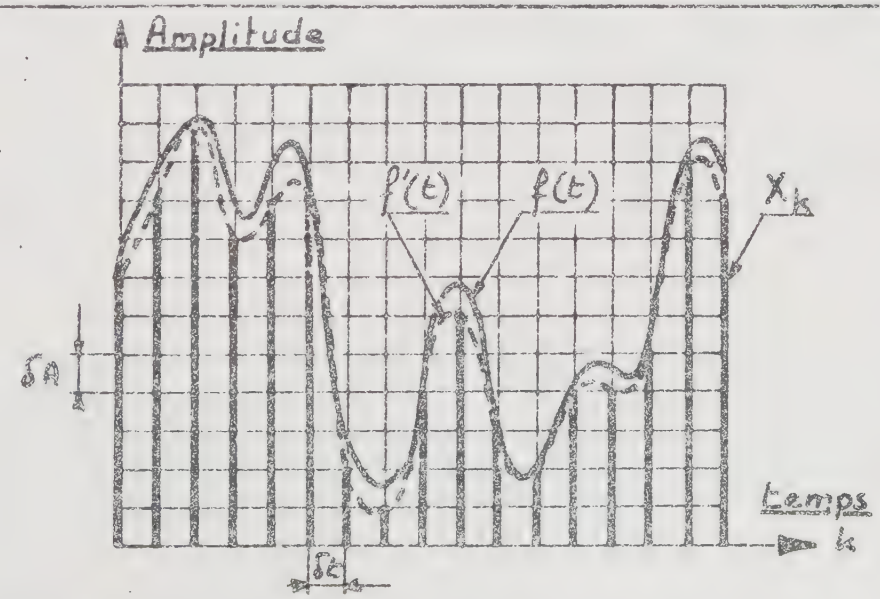


Fig II-8 Origine du bruit de quantification : la courbe $f(t)$ et la courbe $f'(t)$ reconstruite à partir de ses échantillons X_k quantifiés en amplitude ne coïncident pas exactement.

présent. On peut même montrer qu'aucun filtre physiquement réalisable ne peut être un passe-bande parfait (fenêtre spectrale carrée). En effet, la fenêtre temporelle correspondante est de la forme $\frac{\sin t}{t}$ et s'étend de $-\infty$ à $+\infty$. Or un filtre physique ne peut opérer que sur les valeurs présentes ou passées du signal ($-\infty < t \leqslant 0$), donc sa fenêtre temporelle est limitée et il en résulte une fenêtre spectrale "arrondie", sans solution de continuité. La forme du passe-bande idéal peut être approchée d'assez près; mais cela entraîne un allongement prohibitif de la fenêtre temporelle, c'est-à-dire de la "constante du temps" du filtre qui échappe alors à la relation $\Delta t \cdot \Delta F \simeq 1$.

Sans aller plus loin dans la discussion, retenons que la notion de spectre instantané dépend des moyens d'observation du signal; chaque fenêtre temporelle, chaque filtre ou banc de filtres produit un spectre instantané particulier.

II.1.2.4. Loi de KUPFMÜLLER (bib. II-3)

Dans la mesure où un même signal peut être émis avec des durées différentes, sa bande passante ΔF et sa durée Δt sont inversement proportionnelles.

$$\Delta F \cdot \Delta t = \mu \qquad \text{constante de Kupfmüller .}$$

La constante μ caractérise le signal. Cette propriété, très générale, est inscrite dans la transformation de FOURIER. En effet si $F(\omega)$ est la transformée du signal $f(t)$, un signal $f(at)$ aura une transformée

$$\mathcal{F}[f(at)] = \int_{-\infty}^{+\infty} f(at)\, e^{-j\omega t}\, dt = \frac{1}{a} F\left(\frac{\omega}{a}\right)$$

Autrement dit une dilatation de l'échelle de temps entraîne une compression correspondante de l'échelle de fréquence. Pour les signaux impulsionnels usuels, μ est compris entre 0,5 (courbe de Gauss) et 1,5 (signal carré). Ceci permet d'avoir une idée approximative de la bande passante ΔF requise pour décrire un signal de durée Δt dans le domaine fréquentiel.

Une relation analogue caractérise les filtres résonateurs. Δt est alors le "temps de montée" du filtre, et ΔF sa bande passante :

$$\Delta t \cdot \Delta F \simeq 1$$

Ainsi il faudra environ 10 ms pour qu'un signal de 1000 Hz, appliqué brusquement à un filtre de fréquences de coupure 950 et 1050 Hz ($\Delta F = 100$ Hz), arrive à peu près à son régime permanent. Notons que Δt mesure aussi, de manière approximative, la largeur θ de la fenêtre temporelle associée au filtre, dans la mesure où celui-ci est assimilable à un résonateur.

...../

II.1.3. Problèmes liés au traitement numérique.

II.1.3.1. Théorème des échantillonnages

Toutes les opérations de traitement de signal, telles que l'analyse de FOURIER, la convolution, le filtrage, l'intégration, la dérivation, peuvent être effectuées avec des appareillages digitaux. Le théorème des échantillonnages permet en effet d'assimiler un signal continu à une suite de valeurs numériques discrètes. Il s'énonce de la manière suivante (dans le domaine temporel) :

Tout signal $f(t)$ dont la bande est limitée, qui ne possède pas de composantes au delà de la fréquence B (Hertz) est entièrement déterminé par la suite de ses échantillons pris selon un intervalle δt uniforme, inférieur à $\frac{1}{2B}$ seconde. La valeur limite $\delta t = \frac{1}{2B}$ est appelée intervalle de Nyquist.

On peut reconstituer $f(t)$ à partir de la suite de ses échantillons $X(k)$, par une interpolation en $\frac{\sin x}{x}$:

$$f(t) = \sum_{k=-\infty}^{+\infty} X(k) \; \frac{\sin \pi \left(\frac{t}{\delta t} - k\right)}{\pi \left(\frac{t}{\delta t} - k\right)}$$

Pour décrire sans ambiguïté un signal dont la bande est limitée à 100 - 10 000 Hz, par exemple, il faudra disposer au minimum de 20 000 échantillons, prélevés toutes les 50 microsecondes.

Pratiquement il est prudent de conserver une marge de sécurité, et d'échantillonner selon un intervalle $\delta t = \frac{1}{2{,}5B}$ ou $\frac{1}{3B}$, car les signaux usuels ont rarement une bande limitée de manière nette.

Notons que la quantification en temps s'accompagne généralement d'une quantification en amplitude : l'amplitude de chaque échantillon est mesurée avec une précision plus ou moins grande. Dans ce cas le signal $f(t)$ reconstruit à partir de ses échantillons quantifiés représente une approximation plus ou moins fine du signal original $f(t)$ (fig. II-8). Il en résulte une distorsion non linéaire appelée bruit de quantification, dont les fréquences dominantes sont voisines de la fréquence d'échantillonnage.

II.1.3.2. Transformée discrète de FOURIER (DFT) (bib. II-8, II-9)

Si le signal $f(t)$ est représenté par la suite de N échantillons $X(k)$ ($k = 0,1,\ldots\ldots N-1$), sa transformée de FOURIER $F(\omega)$ se présente également sous forme d'une suite de N valeurs :

$$A(r) = \sum_{k=0}^{N-1} X(k)\, e^{-2\pi j r k / N} \qquad r = 0,1,2, \ldots N-1$$

...../

L'intégrale de FOURIER est devenue une somme discrète. Le spectre est défini par une suite de valeurs $A(r)$, où r est l'indice fréquentiel. L'intervalle de quantification fréquentielle δF (Hz) est directement lié à l'intervalle de quantification temporelle δt (secondes) par $N. \delta F . \delta t = 1$

Si, par exemple, nous analysons de cette manière un signal durant 0,0128 secondes, échantillonné 10 000 fois par seconde, ce signal est décrit par la suite des 128 échantillons $X_0 , X_1 ..... X_{127}$. L'intervalle δt vaut $\delta t = 10^{-4}$ s . Le spectre (complexe) est constitué des 128 valeurs

$$A_0 , A_1 ..... A_{127}$$

échelonnées en fréquence avec un intervalle $\delta F = \frac{1}{N. \delta t} = 78,2$ Hz .

Si la fonction $f(t)$ est réelle (cas général en acoustique), le module $|A_r|$ du spectre est une fonction paire c'est-à-dire que (fig II-9a) :

$$|A_0| = |A_{127}|$$

$$|A_1| = |A_{126}|$$

$$|A_{63}| = |A_{64}|$$

Donc si nous ne nous intéressons qu'au module (spectre d'amplitude) ou à son carré (spectre de puissance), nous obtiendrons 64 valeurs du spectre, échelonnées entre 39,1 Hz ($r = 64$) et 4960, 9 Hz ($r = 127$). Ceci est conforme au théorème des échantillonnages : si la bande passante B du signal est limitée à 5000 Hz, alors $\delta t = 10^{-4}$s est l'intervalle de Nyquist, et le signal $f(t)$ est défini sans ambiguïté par les 64 valeurs complexes de $A(r)$, comprises entre 0 et 5 000 Hz.

Notons que, si l'on donne à r des valeurs extérieures à l'intervalle $[0, N-1]$, on retrouve les mêmes valeurs, modulo N . En effet ,

$$\exp\left[- 2\pi j (r + N) k / N\right] = \exp\left[- 2\pi j r k /N\right] . \exp\left[- 2\pi j k\right]$$

Or k est entier ; donc $\exp\left[- 2\pi j k\right] = 1$

Donc la représentation de $f(t)$ par la suite de ses échantillons a pour effet de créer une infinité de spectres images (fig. II-9 b), qui n'apportent naturellement aucune information supplémentaire, mais dont il faut tenir compte, en particulier lorsque l'intervalle δt a été mal choisi (supérieur à l'intervalle de Nyquist). Dans ce cas, les spectres images se chevauchent (fig. II-9 c), et le spectre principal obtenu entre

$$\frac{\omega_1}{2\pi} = 0 \qquad \text{et} \qquad \frac{\omega_2}{2\pi} = \frac{1}{2 \delta t}$$

est entaché d'erreur. Nous verrons au chap. VI un exemple de ce phénomène, appelé "repliement du spectre".

..../

II.1.3.3. Transformée rapide de FOURIER (FFT) (bib. II-8, II-9).

La FFT (Fast Fourier Transform) est une méthode mathématique (COOLEY-TUCKEY) de calcul de la DFT, particulièrement adaptée au calcul numérique. Elle procède par dichotomie successive de la série des N échantillons X_k ("décimation in time "), et permet de réduire considérablement le temps de calcul par rapport à la méthode directe, et ce d'autant plus que les échantillons à traiter sont plus nombreux. En voici le principe.

La série X_k est décomposée en deux séries, comprenant l'une (Y_k) les échantillons d'indice pair, l'autre (Z_k) les échantillons d'indice impair :

$$\left.\begin{aligned} Y_k &= X_{2k} \\ Z_k &= X_{2k+1} \end{aligned}\right) \quad k = 0, 1, 2, \ldots \left(\frac{N}{2} - 1\right)$$

Ces deux séries ont des DFT :

$$\left.\begin{aligned} B_r &= \sum_{k=0}^{\frac{N}{2}-1} Y_k \exp\left(-4\pi j rk/N\right) \\ C_r &= \sum_{k=0}^{\frac{N}{2}-1} Z_k \exp\left(-4\pi j rk/N\right) \end{aligned}\right\} \quad r = 0, 1, 2, \ldots \frac{N}{2} - 1$$

La DFT cherchée est (en décomposant la série X_k)

$$A_r = \sum_{k=0}^{\frac{N}{2}-1} \left\{ X_{2k} \exp\left(-4\pi j rk/N\right) + X_{2k+1} \exp\left[-2\pi j r(2k+1)/N\right] \right\}$$

ici $r = 0,1, \ldots\ldots N-1$ car A_r possède N termes significatifs,

ou encore :

$$A_r = \sum_{k=0}^{\frac{N}{2}-1} Y_k \exp\left(-4\pi j rk/N\right) + \left[\exp\left(-2\pi j r/N\right)\right] \sum_{k=0}^{\frac{N}{2}-1} Z_k \exp\left(-4\pi j rk/N\right)$$

qui s'écrit aussi $A_r = B_r + C_r \exp\left(-2\pi j r/N\right)$ (1)

avec cette fois $r = 0, 1 \ldots\ldots \left(\frac{N}{2} - 1\right)$ car B_r et C_r n'ont que $\frac{N}{2}$ termes significatifs.

Pour $r \geqslant \frac{N}{2}$, les DFT B_r et C_r répètent périodiquement leurs valeurs obtenues pour $0 \leqslant r \leqslant \frac{N}{2}$: on peut remplacer r par $\left(r + \frac{N}{2}\right)$:

$$A_{r+\frac{N}{2}} = B_r + C_r \exp\left[-2\pi j \left(r + \frac{N}{2}\right) / N \right]$$

ou $$A_{r+\frac{N}{2}} = B_r - C_r \exp\left(-2\pi r j / N \right) \qquad r = 0, 1, \ldots\left(\frac{N}{2} - 1\right) \quad (2)$$

...../

Donc l'expression (1) permet de calculer les premiers $\frac{N}{2}$ points de la DFT de X_k , et l'expression (2) permet de calculer les $\frac{N}{2}$ points restants. Or la DFT se calcule en un temps proportionnel à N^2 ; donc le calcul de A_r et B_r demande un temps proportionnel à $2 \left(\frac{N}{2} \right)^2 = \frac{N^2}{2}$, plus N opérations pour recombiner les A_r .

L'intérêt majeur de la méthode est sa récurrence : la même méthode s'applique au calcul de B_r et C_r , tant que le nombre d'échantillons à traiter est multiple de 2.

Si l'on a choisi un nombre N égal à une puissance de 2, soit $N = 2^n$, on peut faire n réductions semblables. On aboutit à une fonction en deux points. La DFT d'une fonction à 1 point est naturellement l'échantillon lui-même; tout le calcul est ainsi réduit à environ $2 n N$ opérations arithmétiques, au lieu des KN^2 opérations du calcul direct.

A titre comparatif, le traitement par FFT de $8192 = 2^{13}$ échantillons requiert 5 secondes de calcul sur IBM 7094, au lieu de 30 minutes par le calcul direct. Pour $N = 128$ le gain de temps, moins considérable, est malgré tout de l'ordre de 10. Sans FFT, l'étude décrite au chapitre VI aurait dû prendre une autre forme.

II.2. LA REPRESENTATION TEMPS-FREQUENCE

II.2.1. Les trois dimensions d'un signal sonore

Si, à chaque instant, on associe au signal son spectre instantané, obtenu au moyen d'un appareillage donné, on peut définir une surface dans l'espace (temps, fréquence, densité d'amplitude ou de puissance). On donne quelquefois le nom "d'objet sonore" (MOLES) au volume compris entre cette surface et le plan de base (fig. II-10). Notons que si la 3° dimension est une densité spectrale de puissance $S(\omega)$ (en W.s), l'aire d'une section $\int_0^\infty S(\omega,t)\, d\omega$ est proportionnelle à la puissance du signal à l'instant t considéré, estimée sur le délai θ de la fenêtre temporelle, et le volume compris entre la surface et le plan de base (temps-fréquence) est proportionnel à l'énergie totale du signal.

Une représentation spatiale est peu utilisable, sinon à titre de démonstration. Mais sa projection normale sur le plan de base donne un document particulièrement pertinent en acoustique perceptive (sonagramme).

Si l'objet sonore était taillé dans un matériau légèrement opaque et disposé dans un faisceau lumineux parallèle et vertical, le sonagramme serait obtenu sur une émulsion photographique disposée dans le plan de base (fig. II-11). Ce procédé hypothétique respecterait même la notation de l'intensité sonore en décibels, car le noircissement obéit à une loi logarithmique en fonction de la quantité de lumière reçue.

Ce document ne permet pas de mesurer la puissance spectrale avec précision. Nous verrons que cela est en accord avec la perception. Néanmoins la troisième dimension peut être traduite par les "lignes de niveau" de l'objet sonore (sonagramme

..../

"contour" fig. II-12). Répétons que le spectre instantané est fonction de la largeur β et de la forme de la fenêtre temporelle utilisée. Ces paramètres sont déterminés par l'appareillage d'analyse, au choix de l'expérimentateur. Mais il serait simpliste de réduire ces paramètres à la seule bande passante du filtre, ou à la seule largeur de la fenêtre temporelle.

II.2.2. Le Sonagraphe (bib. I-2, II-6, II-10)

Dans cet appareil un signal est enregistré en boucle pendant 2,4 secondes sur une piste magnétique, puis répété un certain nombre de fois pour l'analyse (fig. II-13). Un procédé hétérodyne permet de n'utiliser qu'un seul filtre pour couvrir tout le spectre. Rappelons que l'hétérodynage consiste à multiplier le signal à étudier $f(t)$ par une fonction sinusoïdale de pulsation Ω. Le spectre $F(\omega)$ subit alors une simple translation en fréquence, de valeur Ω (fig. II-14). En effet :

$$\mathcal{F}\left[f(t).e^{j\Omega t}\right] = \int_{-\infty}^{+\infty} f(t)\, e^{-j\omega t} e^{j\Omega t} = F(\omega-\Omega)$$

Pour explorer la partie du spectre comprise entre les fréquence $\frac{\omega_1}{2\pi}$ et $\frac{\omega_2}{2\pi}$, on module le signal $f(t)$ avec une tension sinusoïdale de fréquence $\frac{\Omega}{2\pi}$, variable progressivement entre $\frac{\Omega_0+\omega_1}{2\pi}$ et $\frac{\Omega_0+\omega_2}{2\pi}$; $F_0 = \frac{\Omega_0}{2\pi}$ est la fréquence centrale du filtre d'analyse, de largeur de bande ΔF. La fréquence de modulation est commandée dans le sonagraphe par une liaison mécanique : à chaque tour du tambour, la fréquence de modulation augmente de δF, intervalle nécessairement plus petit que ΔF, si l'on veut avoir une définition correcte du spectre.

L'intervalle de fréquence utile est exploré en 270 tours de tambour (environ). Si $\frac{\omega_1}{2\pi} = 0$ et $\frac{\omega_2}{2\pi} = 6000$ Hz, l'intervalle δF entre deux traits successifs ("pas" d'analyse) vaut environ 22 Hz. Cette valeur n'a pas été choisie au hasard : le filtre "étroit" du sonagraphe a une largeur de bande $\Delta F = 45$ Hz, donc la raie la plus fine que l'on puisse trouver dans le spectre $S(\omega)$ du signal est décrite avec au moins 2 traits sur le sonagramme.

Du point de vue de l'expérimentateur, l'intérêt du sonagraphe est de permettre une quantité de réglages, y compris ceux qui ne sont pas prévus par le constructeur. Ainsi la courbe de réponse de l'appareil peut être modifiée en agissant manuellement sur l'amplificateur de marquage, au cours de l'analyse. On peut également comprimer ou dilater temps et fréquence indépendamment en modifiant les vitesses d'enregistrement et de reproduction, et en combinant ces modifications avec une transposition extérieure du message. On peut encore utiliser des tambours de différents diamètres, modifier artificiellement la dynamique du marquage en superposant plusieur marquages sur le même document, etc...

Dans ce sens le Sonagraphe semble être plus un appareil de codage visuel qu'un appareil d'analyse au sens physique du terme. Cependant la latitude laissée à l'expérimentateur porte essentiellement sur la dimension d'amplitude spectrale, alors que les mesures de temps et de fréquence sont toujours reproductibles.

...../

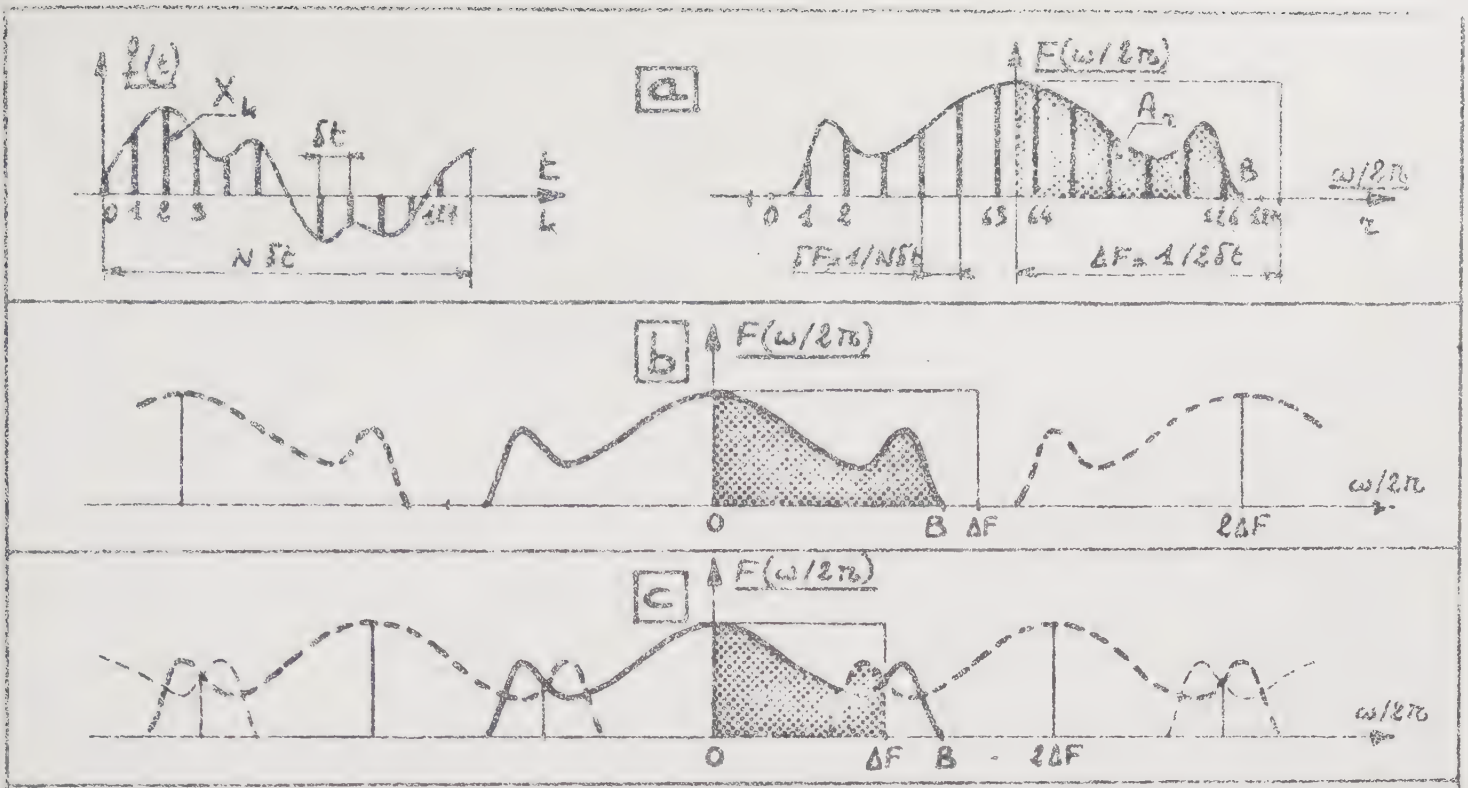


Fig II-9 Transformée discrète de FOURIER. a) Echantillonnage en temps et en fréquence. b) Spectres-images, dans le cas d'un échantillonnage correct ($\delta t \leq 1/2B$). c) Repliement du spectre, dans le cas d'un échantillonnage incorrect ($\delta t > 1/2B$).

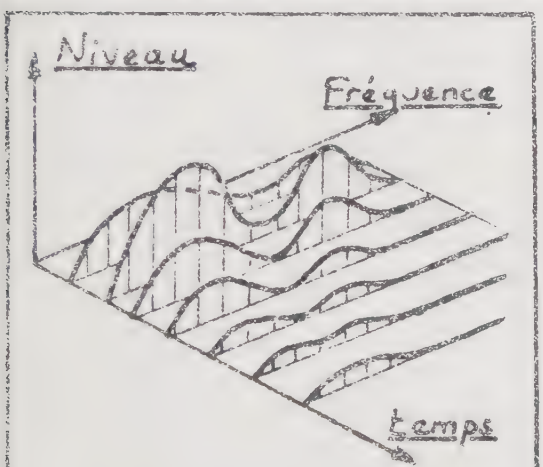


Fig II-10 Objet sonore, représenté par une suite de spectres instantanés.

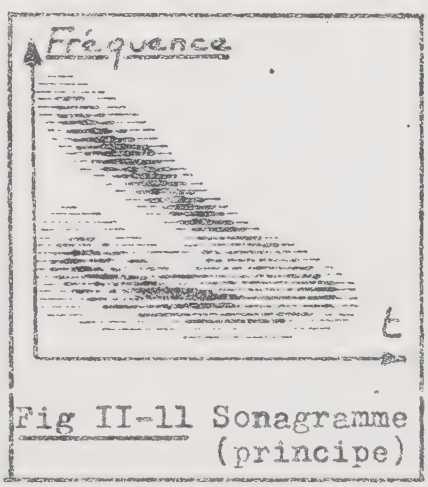


Fig II-11 Sonagramme (principe)

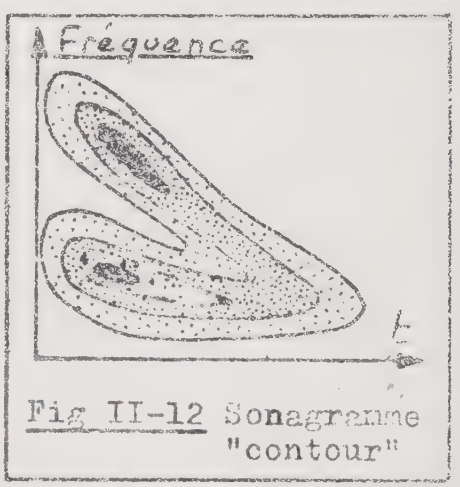


Fig II-12 Sonagramme "contour"

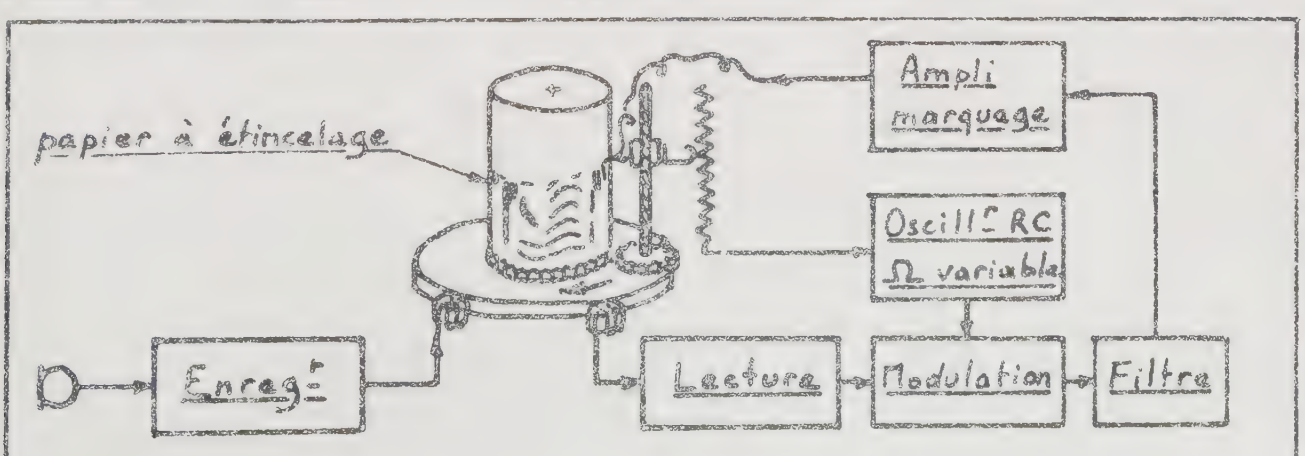


Fig II-13 Principe du Sonagraphe

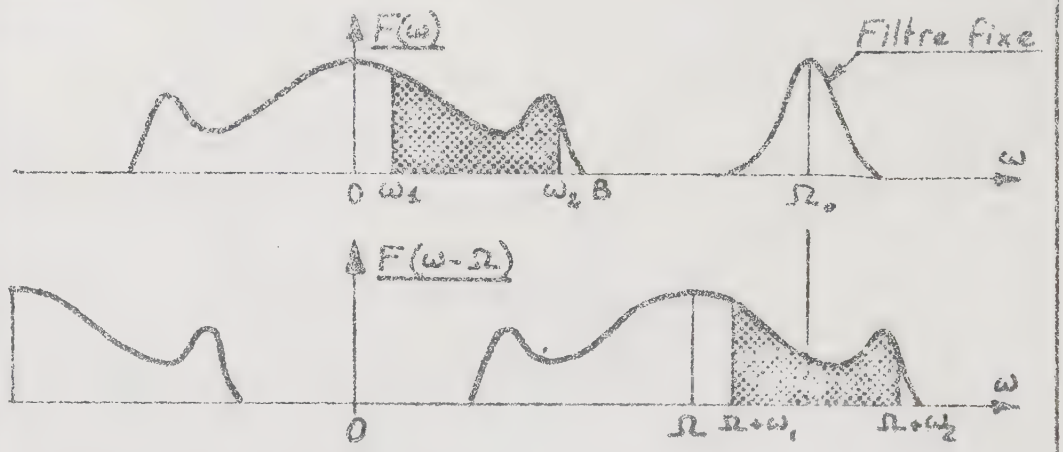


Fig II-14 Principe de la modulation hétérodyne : le signal $f(t)$ est multiplié par la fonction $\sin \Omega t$, de pulsation Ω variable, de façon à pouvoir utiliser le filtre fixe, de fréquence centrale $\Omega_0 / 2\pi > B$.

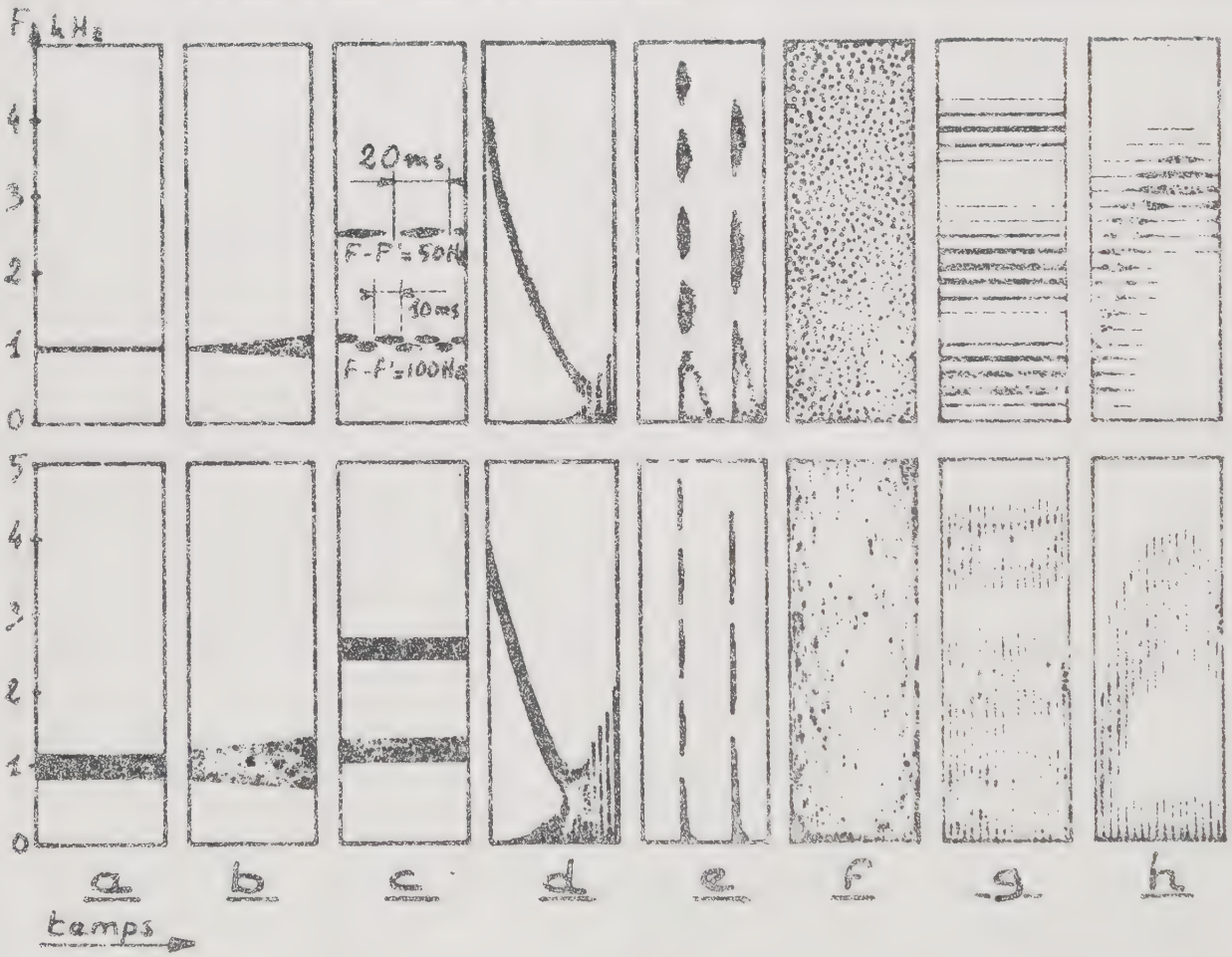


Fig II-15 Sonagrammes schématiques de quelques sons élémentaires (en haut analyse en bande étroite de 45 Hz ; en bas analyse en bande large de 300 Hz).
a) Son simple (1000 Hz). b) Son simple d'amplitude croissante. c) Sons de fréquences voisines.
d) Son simple de fréquence décroissante. e) Impulsions
f) Bruit blanc. g) Son harmonique avec formants fixes.
h) Formant évolutif.

En bref le Sonagraphe est un analyseur d'une grande souplesse. Celle-ci est nécessaire dans les études concernant la parole, et nous devrons donner à nos programmes une souplesse comparable si nous voulons traiter les mêmes problèmes avec une machine à calculer.

II.2.3. Représentation temps-fréquence de quelques sons élémentaires.

Sur la figure II-15 se trouvent réunis quelques sonagrammes schématiques, illustrant les considérations précédentes.

a) Un son simple (sinusoïdal) de fréquence 1000 Hz est représenté par un trait horizontal. Ce trait est fin en bande étroite (ΔF = 45 Hz), épais en bande large (ΔF = 300 Hz).

b) Si le niveau sonore croît, le noircissement croît d'abord, puis le trait s'épaissit, à mesure que les filtres connexes sont concernés. Ceci est dû à la forme de la fenêtre spectrale (courbe de réponse du filtre), dont les pentes sont de l'ordre de quelques dizaines de décibels par octave. Il est vraisemblable, comme nous le verrons plus loin, qu'un processus analogue existe dans l'oreille.

c) Cette diffusion du tracé pourrait donner à penser que deux sons périodiques de fréquences voisines ne peuvent être discernés. Il n'en est rien, car ces deux sons produisent un battement dont la fréquence est parfaitement mesurable.

d) Un son simple (pseudo-sinusoïdal) de fréquence décroissante, donne un trait descendant. Lorsque la fréquence devient égale ou inférieure à la bande passante du filtre, c'est-à-dire lorsque la fenêtre temporelle devient plus petite que la période, le spectre instantané devient celui d'impulsions qui se succèdent en synchronisme avec les périodes du signal. Ce phénomène est plus sensible en bande large; il permet de mesurer la fréquence du signal par l'intermédiaire de sa période.

e) Une impulsion est représentée par un trait vertical, suivi d'un traînage plus ou moins long qui dépend, comme le bord d'attaque, de la réponse temporelle du filtre. On détecte facilement sur le sonagramme les zéros de la transformée de FOURIER de l'impulsion; ceux-ci diffèrent pour chaque forme d'impulsion. Mais il est impossible de reconstruire précisément la forme temporelle du signal puisque le spectre de phase a disparu.

f) Un bruit blanc ou coloré se traduit par un grisé plus ou moins uniforme dans les deux modes d'analyse. Néanmoins le grain du tracé est différent selon la largeur d'analyse.

g) Un son harmonique est représenté par un réseau de raies équidistantes. Chaque harmonique possède son amplitude propre. Si la fréquence fondamentale F_o est comparable à la largeur de bande ΔF, les raies fusionnent. Si F_o est inférieure à ΔF, la période du fondamental se traduit comme en d) par l'intervalle entre traits verticaux successifs.

h) On trouve souvent dans le spectre certaines zones fréquentielles favorisées nommées formants, dont l'évolution est distincte de l'évolution du son harmonique ou du bruit qui les supportent. La notion de formant est pratique, mais souvent ambiguë, comme nous le verrons au chapitre IV.

...../

II.2.4. Représentation temps-fréquence de la parole

La figure II-16 comporte plusieurs représentations de la phrase "les petits oiseaux chantent", prononcée par un locuteur masculin. L'analyse a été effectuée selon diverses largeurs de bande, allant de 10 à 300 Hz.

Avec $\Delta F = 10$ Hz la constante de temps Δt du filtre est trop grande (de l'ordre de 100 ms) par rapport au son étudié, si bien que la résolution fréquentielle obtenue ne sert à rien. La bande $\Delta F = 45$ Hz permet de bien discerner les composantes harmoniques d'une voix dont le fondamental est supérieur à environ 90 Hz (deux fois la bande passante d'analyse), mais les phénomènes très brefs sont estompés. L'analyse avec $\Delta F = 300$ Hz met en évidence les formants et les phénomènes brefs, en intégrant l'énergie des harmoniques voisins, pour les voix dont le fondamental est inférieur à 300 ou 400 Hz. L'analyse avec $\Delta F = 150$ Hz est intermédiaire : l'aspect harmonique est perdu, l'aspect formantique n'est pas encore très net, sauf pour les fondamentaux graves.

La représentation en "contour" avec $\Delta F = 300$ Hz apporte une information complémentaire sur l'amplitude spectrale, puisque les courbes de niveau sont tracées à intervalles de 6 dB(dynamique totale 36 dB); mais c'est au détriment de la résolution temporelle puisque ce tracé nécessite une intégration portant sur 5 à 10 ms. La structure fine du son disparaît; seule subsiste l'enveloppe de l'objet sonore.

On adopte l'une ou l'autre de ces représentations suivant l'information recherchée, ou même leur superposition. L'oreille est capable d'apprécier les plus fins détails visibles sur ces documents et, réciproquement, il n'est guère de phénomène audible que l'on ne puisse déceler avec le Sonagraphe. On est ainsi conduit à penser que l'hypothétique analyseur auditif, même s'il fonctionne de tout autre façon, est capable de transformer un signal en une suite de spectres instantanés, par l'usage d'une fenêtre temporelle de largeur comprise entre $\frac{1}{300} = 3$ ms et $\frac{1}{10} = 100$ ms . Cette remarque est évidemment simpliste, et le problème sera approfondi au chapitre III.

II.3. INFORMATION, CODAGE ET REDONDANCE (bib. II-3, II-4, II-11)

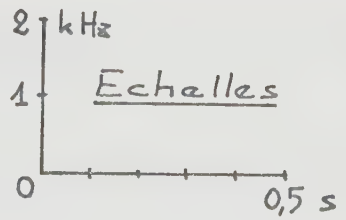
La Théorie de l'Information permet d'étudier quantitativement les messages de toutes natures et leur transmission. Elle donne une base mathématique au schéma de la communication entre les systèmes; les notions fondamentales peuvent être appliquées à l'étude de la communication entre individus, moyennant certaines précautions. Nous exposerons de manière élémentaire les fondements de la Théorie de l'Information, en nous limitant tout d'abord au domaine des systèmes physiques, où elle n'est pas contestable. Puis nous chercherons à montrer l'intérêt et les limites de son application dans le domaine de la communication entre individus.

II.3.1. Mesure de l'information

Lorsqu'un message est constitué d'une suite de n éléments distincts et identifiables, choisis dans un répertoire de N éléments de référence dont les probabilités d'apparition sont $p(i)$, on peut associer à ce message une quantité d'information H, déterminée par la formule de WIENER-SHANNON :

...../

Fig II-16 Représentation temps-fréquence de la phrase "les petits oiseaux chantent" , selon diverses largeurs d'analyse.



$$H = -n \sum_{i=1}^{N} p(i) \log_{\varepsilon} p(i)$$

avec bien sûr $\sum_{i=1}^{N} p(i) = 1$

Lorsque tous les symboles sont équiprobables ($p(i) = p_e \ \forall \ i$) l'information s'écrit plus simplement

$$H = - n \log_2 p_e$$

H s'exprime en bits (bit = binary unit). C'est le nombre de choix binaires qui définissent le message sans ambiguïté. La quantité $S = \frac{H}{n}$, information moyenne par symbole, qui ne dépend que des probabilités $p_{(i)}$, est appelée par SHANNON entropie négative, ou négentropie, par analogie avec la thermodynamique. C'est aussi la longueur moyenne L des symboles utilisés lorsqu'ils sont représentés par des nombres binaires. Prenons quelques exemples :

a) Message constitué d'une suite de 10 symboles binaires (binits, abréviation de binary digits) équiprobables

p.ex. : 1 0 1 1 0 0 0 1 0 1

Chaque symbole, 0 ou 1, a la même probabilité ($p_o = p_1 = 0,5$) estimée à partir d'un grand nombre de messages du même type.

L'information H_1 du message vaut

$$H_a = - 10 \log_2 0,5 = 10 \log_2 2 = 10 \text{ bits}$$

Autrement dit, l'entropie S_a du code utilisé vaut 1 bit.

b) Message constitué d'une suite de 10 chiffres équiprobables

p.ex.: 1 5 6 3 2 9 5 8 7 0

Chaque symbole, choisi de zéro à neuf, possède une probabilité $p_{(i)} = 0,1$. L'information H_b est alors

$$H_b = - 10 \times 3,32 \log_{10} 0,1$$

$$= 33,2 \text{ bits}$$

ou $S_b = 3,32$ bits

On voit qu'un chiffre décimal contient en moyenne 3,32 fois plus d'information qu'un chiffre binaire. Pour acheminer le même message avec des symboles binaires équiprobables, il faudra 34 symboles au lieu de 10.

Ces deux exemples permettent d'introduire la notion de code, binaire dans l'un, décimal dans l'autre. Le second message pourrait même s'exprimer par un seul

...../

symbole dans un code comprenant un répertoire de 10^{10} symboles différents : plus le répertoire est étendu, plus le message peut s'exprimer sous une forme succincte.

Notons que tous les codages d'un même message ne sont pas exactement équivalents : c'est ainsi que, pour coder un nombre entier compris entre 0 et 99, par exemple, on peut utiliser soit 2 symboles décimaux, soit 7 symboles binaires. Or, le message est choisi parmi 100 possibilités. Sa probabilité est 0,01 et l'information qui lui est associée vaut :

$$H = - 1 \times (100 \times 0{,}01 \log_2 0{,}01)$$

$$H = 3{,}32 \log_{10} 100 = 6{,}64 \text{ bits}$$

L'information transportée par 2 symboles décimaux vaut

$$H' = - 2 \log_2 0{,}1 = 6{,}64 \text{ bits}$$

L'information transportée par 7 symboles binaires vaut :

$$H'' = - 7 \log_2 0{,}5 = 7 \text{ bits}$$

Dans ce cas, l'utilisation de symboles binaires constitue une perte d'efficacité : le codage binaire n'est pas optimum pour le message considéré.

L'utilisation de symboles non équiprobables mène également à une perte d'efficacité du codage. Supposons que dans un certain type de message binaire, on trouve en moyenne trois fois plus de "un" que de "zéro". Les probabilités sont alors :

$$p_1 = 0{,}75$$

$$p_o = 0{,}25$$

A un message constitué comme précédemment (ex.a) de 10 symboles binaires correspond alors une information H_c

$$H_c = - 10 (0{,}25 \log_2 0{,}25 + 0{,}75 \log_2 0{,}75)$$

$$= 33{,}2 (0{,}25 \log_{10} 4 + 0{,}75 \log_{10} \frac{4}{3})$$

$$= 33{,}2 \times 0{,}24 = 8{,}1 \text{ bits}$$

ou $S_c = 0{,}81$ bit .

II.3.2. Efficacité du codage - Redondance

Le premier théorème de SHANNON, relatif à la transmission sans bruit parasite (ou sans erreur), indique qu'un message codé avec N symboles distincts d'entropie S_m (entropie de source) peut être transmis au moyen d'un autre code à N symboles, pourvu que l'entropie S du message exprimé dans le nouveau code soit égale ou supérieure à S_m .

...../

De ce théorème, comme des exemples précédents, se dégage l'idée de codage optimum d'un message, réalisé lorsque l'entropie du message exprimé dans le nouveau code est égale à l'entropie de source. En général, le codage optimum conduit à un répertoire très étendu ($N \longrightarrow \infty$); et dans la pratique cela pose des problèmes de décodage et d'identification. Si le codage n'est pas optimum ($S > S_m$), la perte d'efficacité, ou redondance, s'exprime par $R = \frac{S - S_m}{S}$ $\quad 0 \leqslant R \leqslant 1$

La redondance indique, sous forme d'un pourcentage, ce qui est "en trop" dans le message.

Ainsi, un message constitué de symboles binaires dont les probabilités sont 0,75 et 0,25 (entropie de source 0,81 bit) peut être transmis au moyen d'un code décimal équiprobable (longueur moyenne L = 3,32 bits), et ceci avec une redondance

$$R = \frac{3,32 - 0,81}{3,32} = 0,76$$

La notion de redondance est fondamentale; nous verrons qu'elle est liée à l'indestructibilité des messages. Auparavant, il nous faut évoquer le cas où les symboles du code ne sont pas indépendants les uns des autres.

II.3.3. Cas où les symboles ne sont pas indépendants (bib. II-13)

Si on a remarqué qu'un symbole (k) suivait souvent un symbole (i), on peut attribuer au digramme (i, k) une probabilité p (k/i), qui est la probabilité de voir le symbole (k) apparaître quand le symbole (i) est apparu. On peut établir ainsi une matrice de N^2 termes pour un répertoire de N symboles, étant entendu qu'il existe N relations analogues à

$$\sum_{k=1}^{N} p\,(k/i) = 1 \qquad\qquad i = 1,2, \ldots N$$

Rappelons que la probabilité conditionnelle p (i/k) peut être calculée à partir de la loi de BAYES

$$p\,(i/k) = \frac{p\,(i) \cdot p\,(k/i)}{\sum_{i=1}^{N} p\,(i) \cdot p(k/i)}$$

Le dénominateur n'est autre que la probabilité p (k) du symbole k -

Dans ce cas, l'entropie d'un message exprimé avec ces symboles prend une nouvelle expression

$$S = -\sum_{i=1}^{N} p\,(i) \sum_{k=1}^{N} p\,(k/i) \log_2 \left[p\,(k/i) \right]$$

Nous reprendrons les données de l'exemple (c) précédent (p(1) = 0,75 ; p(o) = 0,25), en ajoutant les probabilités conditionnelles suivantes (exemple d) :

p (0/1) = 0,2 $\qquad$ p (1/0) = 0,6

p (1/1) = 0,8 $\qquad$ p (0/0) = 0,4

...../

C'est-à-dire que 1 est plus souvent suivi de 1 que de zéro - Remarquons que, dans cet exemple, il suffit de connaître p (o), p (1) et une seule des 4 probabilités conditionnelles, pour calculer les 3 autres.

Tous calculs faits, nous trouvons une nouvelle entropie $S_d = 0,78$ bit, au lieu de $S_c = 0,81$ lorsque les symboles étaient indépendants. La dépendance introduite par les probabilités conditionnelles a entraîné une redondance

$$R_{dc} = \frac{0,81 - 0,78}{0,81} \simeq 0,04$$

On peut, d'une manière analogue, faire intervenir les probabilités de trigrammes ou de polygrammes d'ordre supérieur, qui ont toujours pour effet de diminuer l'efficacité du codage par rapport à l'équiprobabilité. Les calculs deviennent très vite inextricables lorsque le nombre d'éléments du répertoire augmente.

II.4. CANAL, BRUIT, MESSAGES CONTINUS (bib. II-4)

II.4.1. Canal et sûreté de transmission

La transmission d'un message dans le temps ou dans l'espace se fait par l'intermédiaire d'un certain canal physique (disque, film, ondes hertziennes, etc...) susceptible d'accepter l'information dans un certain code. Il est bien évident que le canal et le message à transmettre doivent être adaptés l'un à l'autre : la transmission d'une onde acoustique (plus exactement de son analogue électrique) par un canal binaire nécessite un codage à la source (PCM : modulation par impulsions codées, par exemple) et un décodage à la réception.

Un canal est caractérisé par la sécurité de la transmission qu'il assure, c'est-à-dire par l'improbabilité des erreurs possibles. Celle-ci conduit à la capacité C_s du canal par symbole. Si, par exemple, en appliquant 1 signe binaire à l'entrée d'un canal binaire on peut toujours recueillir à la sortie ce même signe, la capacité du canal est

$$C_s = 1 \text{ bit par symbole binaire}$$

La capacité du canal peut également s'exprimer en bits par seconde (ou bauds), dans la mesure où l'on connaît le nombre K de symboles transmis par seconde :

$$C = K\, C_s \text{ bits/s}$$

Mais, en général, la transmission (codage et décodage compris) est entachée de bruits parasites ou d'erreurs. La probabilité non nulle de ces erreurs fait que la capacité réelle est inférieure à celle que l'on prévoyait en l'absence de bruit; pour le canal binaire précédent

$$C_s < 1 \text{ bit pour symbole binaire en présence de bruit}$$

Si l'entropie de source est S (information moyenne par symbole) on aura, d'une manière générale $C_s \leqslant S$

...../

Le second théorème de SHANNON, relatif à la transmission d'information en présence de bruit, indique que la transmission peut s'effectuer "sans erreur" (c'est-à-dire avec un taux d'erreur aussi petit que l'on veut), même si la capacité C_s du canal est inférieure à l'entropie de source du message. Il faut pour cela coder le message avec une redondance supérieure à un certain seuil lié à C_s, ce qui a pour effet de diminuer la cadence de transmission de l'information.

Supposons, par exemple, que le canal binaire précédent ait une probabilité d'erreur $p = 10^{-1}$ et que l'entropie de source soit $S = 1$ bit. La capacité calculée du canal est $C_s = 0,55$ bit par symbole binaire. D'après le théorème de SHANNON, la sûreté de transmission peut être améliorée si l'information est transmise à un rythme inférieur à 0,55 bit par symbole binaire, c'est-à-dire avec une redondance

$$R \geqslant \frac{1 - 0,55}{1}$$

$$\text{Soit} \quad R \geqslant 0,45$$

On peut introduire une redondance $R = 0,67$ en répétant deux fois chaque symbole du message-source, ce qui divise par 3 la cadence de transmission du message. On montre que la probabilité d'erreur tombe alors à $p_1 = 3.10^{-2}$.

II.4.2. Cas des messages continus

La notion de capacité d'un canal a été étendue aux messages continus : considérons un canal de largeur de bande B (Hertz), dans lequel s'écoule un signal $f(t)$ dont la puissance quadratique moyenne est P_s (Watt), en présence d'un bruit blanc gaussien de puissance quadratique moyenne P_B (Watt) - La capacité C (bits/seconde) d'un tel canal est donnée par

$$C = B.\log_2 \left(1 + \frac{P_s}{P_B} \right)$$

Nous pouvons ainsi calculer les débits informationnels nécessaires pour transmettre la parole selon les critères de la "haute-fidélité" et du téléphone :

Hi-Fi

La bande passante est de l'ordre de 50 - 15 000 Hz, soit $B = 1,5.10^4$ Hz; la dynamique est de l'ordre de 60 décibels, c'est-à-dire

$$\frac{P_s}{P_B} \simeq 10^6$$

D'où $C \simeq 1,5.10^4 \times 3,32 \log_{10} 10^4 \simeq 2.10^5$ bits/seconde.

Téléphone

La bande passante est limitée à 300 - 3400 Hz, soit $B = 3,1.10^3$; la dyna-

..../

mique est de l'ordre de 20 décibels, c'est-à-dire

$$\frac{P_s}{P_B} \simeq 10^2$$

D'où $C = 3{,}1.10^3 \times 3{,}32 \log_{10} \simeq 2.10^4$ bits/seconde

Nous verrons plus loin que ces valeurs sont considérables eu égard à l'information utile de la parole (quelques dizaines de bits par seconde). Comme le prix des appareillages de télétransmission est approximativement proportionnel au débit informationnel requis, ceci justifie économiquement les recherches multiples qui ont pour but de trouver un codage peu redondant de la parole.

II.5. INFORMATION ET COMMUNICATION ENTRE INDIVIDUS (bib. II-11, II-12, II-13).

Nous avons vu que tout échange d'information impliquait un émetteur, un récepteur et un canal assurant la transmission plus ou moins sûre de l'information sous forme de symboles discrets et codés. La mesure de l'information transmise implique la connaissance des symboles et de leurs probabilités sur l'ensemble des messages possibles. Cependant, la nature même de l'information doit être précisée, ainsi que certains aspects des mécanismes de codage et décodage.

II.5.1. Intersection des répertoires

Dans un système technique, l'émetteur et le récepteur sont étudiés en fonction du code approprié au canal. Le répertoire des symboles est commun, par construction. Dans la communication humaine, au contraire, les codes se sont élaborés en fonction des individus, de façon à permettre l'échange d'information entre le plus grand nombre de membres d'une même société. Il est bien évident que deux individus ne peuvent communiquer que dans la mesure où leurs répertoires personnels ont une partie commune (fig. II-17).

II.5.2. Débit informatif du message parlé

Le message parlé confoie des informations de natures diverses, que nous allons essayer de classer.

II.5.2.1. Information physique

L'information acoustique, au sens physique du terme, est liée à la forme de l'onde acoustique; nous avons vu plus haut que son débit H_1 était de l'ordre de 2.10^5 bits/s. Il semble que le cerveau soit incapable de traiter un tel volume d'information.

...../

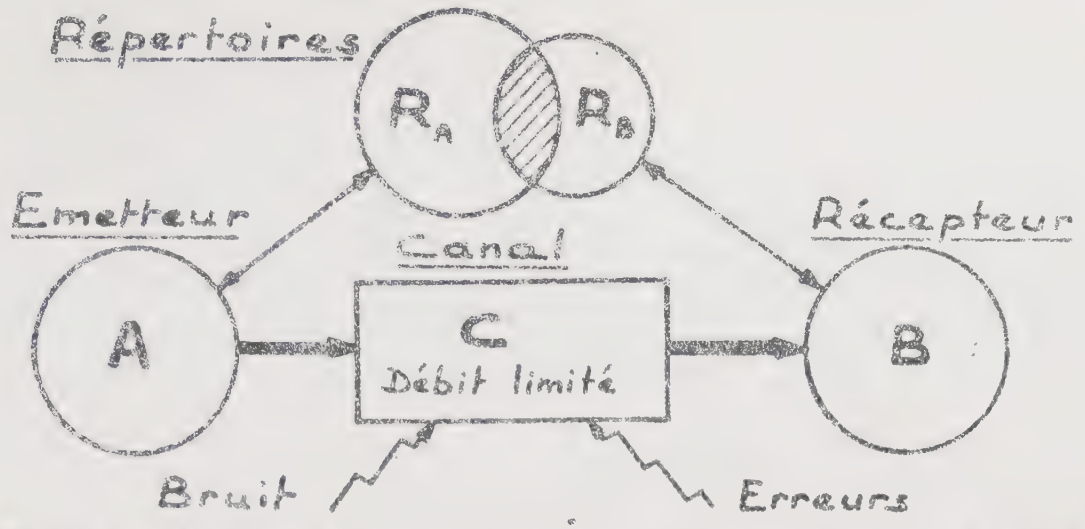


Fig II-17 Schéma de la communication : les individus A et B ne peuvent communiquer, par l'intermédiaire du canal C, que dans la mesure où leurs répertoires R_A et R_B ont une partie commune.

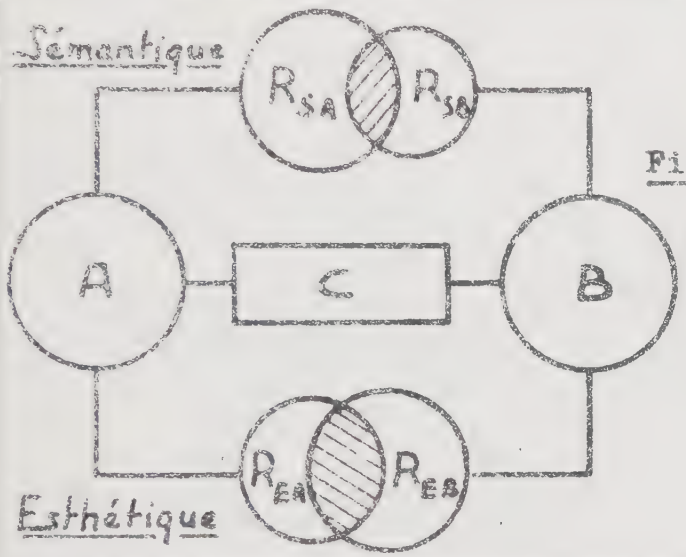


Fig II-18

Le même canal peut acheminer deux types différents d'information, ayant chacun leur propre répertoire (d'après A.MOLES).

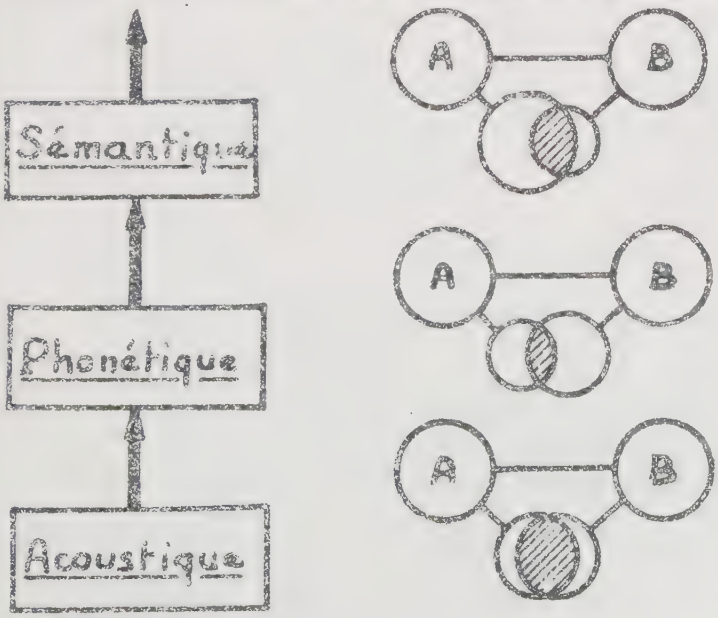


Fig II-19

Hiérarchie des niveaux de communication : en passant d'un niveau au niveau supérieur, les symboles s'organisent en supersymboles, et sont puisés dans des répertoires propres (d'après A.MOLES).

II.5.2.2. Information perceptible

L'information acoustique, au sens perceptif du terme, est certainement plus réduite. Deux messages sont perceptivement identiques s'ils sont indiscernables l'un de l'autre pour une majorité d'observateurs. Il n'est pas nécessaire, par exemple, de transmettre le spectre de phase instantanée; le débit acoustique perceptible H_2 est donc au moins deux fois plus faible que le débit acoustique physique. Nous verrons, au chapitre V, que la parole peut être synthétisée avec une excellente qualité au moyen d'une dizaine de paramètres. Si l'on admet que chacun de ces paramètres, supposé indépendant, peut prendre 1 valeur toutes les 5 millisecondes parmi 100 valeurs équiprobables, on est conduit à un débit

$$H_2 \quad - 10 \times 200 \log_2 10^{-2}$$

$$H_2 \quad 14\ 000 \text{ bits/s}$$

Il ne s'agit là, bien sûr, que d'un ordre de grandeur (certainement majoré, car les paramètres ne sont pas totalement indépendants), qui montre essentiellement la grande redondance du message physique par rapport au message perceptible

$$R_{12} \quad \frac{2.10^5 - 1,4.10^4}{2.10^5} = 0,93$$

II.5.2.3. Information phonétique

Un autre type d'information nous est donné par la phonétique. Les phonéticiens s'accordent à noter le message parlé sous forme d'un nombre limité de symboles appelés phonèmes. S'il est incontestable que la notation phonétique définisse sans ambiguïté le sens du message, les phonèmes ne peuvent être considérés comme des éléments physiquement identifiables (nous reviendrons sur ce point). Nous pouvons cependant admettre qu'il s'agit d'un codage de la parole préservant son intelligibilité. Le nombre de phonèmes utilisés dans une langue donnée est généralement compris entre 20 et 40. Un locuteur utilisant $2^5 = 32$ phonèmes équiprobables, à la cadence de 10 par seconde, fournit un débit d'information phonétique de l'ordre de

$$H_3 = -10 \log_2 2^{-5} = 50 \text{ bits/seconde}$$

En réalité, l'information phonétique est plus faible car, d'une part, les symboles phonétiques ne sont pas équiprobables et, d'autre part, ils sont liés les uns aux autres par des probabilités conditionnelles dont la distribution caractérise la langue (voir chap. IV).

II.5.2.4. Information transportée par les mots

Si l'on considère les mots eux-mêmes comme éléments transmis par le message parlé, le débit d'information est encore plus faible : supposons tout d'abord qu'un locuteur dispose d'un répertoire de 10 000 mots équiprobables, et qu'il parle avec un débit moyen de 2,5 mots par seconde - Son débit est alors

$$H_4 = - 2,5 \log_2 10^{-4} \simeq 34 \text{ bits/s}$$

...../

Cette estimation est bien trop forte; en effet, les contraintes grammaticales et sémantiques, dues au contexte d'une conversation courante, font que l'auditeur doit généralement choisir parmi une dizaine de mots seulement à chaque instant, ce qui conduit à un débit

$$H'_4 = -2{,}5 \log_2 10^{-1} \simeq 8 \text{ bits/s}$$

Précisons bien que ce dernier calcul n'a pas de sens du point de vue strictement mathématique : la mesure de l'information repose sur une base statistique; à chaque symbole doit être associée une probabilité bien définie, établie sur l'ensemble des messages.

Ces probabilités existent à l'échelle de la langue (loi de ZIPF), mais ne concernent pas la communication individuelle, ce qui est tout-à-fait contraire aux hypothèses de la Théorie de l'Information.

Pour approximatifs qu'ils soient, ces calculs fournissent une distinction commode entre deux types d'information, esthétique et sémantique; ils montrent, par ailleurs, une hiérarchie des niveaux de communication et une organisation des répertoires dont la Théorie de l'Information ne peut seule rendre compte.

II.5.3. Information sémantique et information esthétique (bib. II-11, II-12)

La signification du message parlé dépend de l'interprétation de l'individu récepteur, qui réagit selon son conditionnement, son intelligence et les circonstances présentes - Aussi semble-t-il hasardeux de parler de l'information sémantique comme d'une quantité mesurable. Pourtant, le même message, énoncé en français, en anglais, noté en sténographie, en idéogrammes, ou phonétiquement, possède bien un invariant, qui correspond statistiquement à une réaction comparable de la part de groupes sociologiques comparables. En vérité, l'échange d'information sémantique est l'objet même de la communication parlée.

Faute de savoir définir précisément l'information sémantique de la parole, nous admettrons qu'elle est contenue dans l'information phonétique qui, elle, est mesurable, puisqu'elle peut être codée avec un petit nombre de symboles discrets. Donc, lorsque nous parlerons d'information sémantique de manière quantitative, il faudra comprendre information phonétique.

L'information esthétique (MOLES) ou ectosémantique (MEYER-EPPLER), représente une partie de la différence entre l'information perceptible et l'information sémantique. Elle est tout ce qui ne contribue pas directement à l'intelligibilité, ce qui permet d'identifier le locuteur, de connaître son âge, son sexe, son état physique et émotionnel, de percevoir des nuances propres au langage parlé. Cette définition n'est pas très positive, quantitativement parlant; on peut seulement affirmer que l'information esthétique est contenue dans l'information perceptive, et donc que son débit maximum est de l'ordre de 10 000 bits par seconde.

Pour aller plus loin, il faudrait pouvoir définir et isoler précisément les composantes énoncées plus haut, en chercher les éléments, le débit, la redondance, etc... Cela n'est pas notre propos. Retenons que si l'information sémantique est associée aux invariants de la parole, et estimée par excès à une cinquantaine de bits/seconde, l'information esthétique est associée aux variations non significatives du message, aux champs de liberté du locuteur, et son débit est de l'ordre de 10^4 bits/seconde. Nous nous intéresserons surtout dans la suite à l'information phonétique et sémantique; cependant, nous verrons au passage quelques aspects de l'information esthétique;

le problème essentiel de l'analyse de la parole consiste d'ailleurs à séparer l'une de l'autre objectivement.

On peut supposer que l'information esthétique est transmise au moyen d'éléments discrets choisis dans un répertoire commun au locuteur et à l'auditeur, par le même canal que l'information sémantique. La figure II-18 montre le schéma correspondant (MOLES).

Ces tentatives d'application de la Théorie de l'Information à la communication parlée, nous éloignent de plus en plus de l'aspect quantitatif initial. Et pourtant ces schémas sont encore tout-à-fait simplistes, en ce qu'ils ne tiennent aucun compte, par exemple, de l'organisation interne des répertoires.

II.5.4. Hiérarchie des niveaux d'information

II.5.4.1. Gain produit par la transmission des mots

Plaçons-nous dans le cas d'une transmission phonétique, dont les 32 symboles équiprobables sont également connus de l'émetteur et du récepteur. Nous avons vu que le débit requis était $H_3 = 50$ bits/seconde.

Nous pouvons former de nouveaux symboles par association de n symboles phonétiques, : nous obtenons ainsi des "supersymboles" ou mots, dont le nombre N croît très vite avec n :

$$N = \frac{32\ !}{(32-n)!}$$

Les 32 symboles élémentaires permettent de constituer environ 10^{36} mots de longueur comprise entre 1 et 32, de longueur moyenne $\overline{n}$ voisine de 32 , transmis à la cadence d'environ 0,3 mot par seconde, avec le débit $H_3 = 50$ bits/seconde. Un répertoire de 10^{36} mots dépasse évidemment l'entendement humain. Supposons le répertoire limité à 10 000 mots, de longueur moyenne $n = 4$ symboles phonétiques, prononcés à la cadence d'environ 2,5 par seconde, soit avec un débit de l'ordre de

$$H_3' = 2{,}5 \log_2 10^4 \simeq 34 \text{ bits/seconde.}$$

Ainsi la limitation du vocabulaire et de la longueur moyenne des mots utilisés s'accompagne-t-elle d'une réduction notable du débit d'information. L'association des symboles phonétiques entre eux pour constituer des mots peut être considérée comme l'application de probabilités conditionnelles très compliquées, qui introduisent une redondance supplémentaire en même temps qu'une plus grande sécurité de transmission. La transmission des mots est beaucoup plus redondante que celle des symboles phonétiques. Si le canal de transmission ne permet qu'un débit de 40 bits/secondes, par exemple, il est possible de transmettre les mots sans erreur en temps réel, mais pas les phonèmes.

Cette remarque sera de la plus grande utilité lorsqu'il s'agira de reconnaissance automatique : la reconnaissance de mots choisis dans un répertoire limité est beaucoup plus simple à mettre en oeuvre que la reconnaissance d'éléments phonétiques.

...../

II.5.4.2. Contraintes syntaxiques

Un processus analogue intervient dans l'association des mots, qui sont répartis dans différentes catégories syntaxiques. Bien que les longueurs et les types de phrases semblent d'une variété infinie, on a pu estimer à une trentaine (MOLES) les modèles de phrases effectivement utilisés dans le langage courant. La reconnaissance des premiers mots suffit à déterminer le type de la phrase, et à compenser d'éventuelles lacunes ou ambiguïtés phonétiques.

L'organisation du message en symboles et supersymboles conduit à distinguer des niveaux d'information hiérarchisés. L'information sémantique, par exemple, peut être examinée au niveau des éléments acoustiques perceptifs, au niveau des éléments phonétiques, au niveau des mots, puis des phrases (fig. II-19). Chaque niveau englobe ceux qui lui sont inférieurs et correspond à un gain dans la sécurité de transmission du message.

II.5.5. Influence du contexte et de la structuration sémantique du message

C'est ici qu'achoppe véritablement la Théorie mathématique de l'Information. La communication entre individus ne se fait pratiquement jamais sans contexte. La probabilité d'un mot ou d'une idée n'est pas une donnée statistique fixée une fois pour toutes dans le répertoire de chacun : elle est fonction des circonstances présentes et du message lui-même.

a) fonction des circonstances

Soient par exemple deux personnes A et B possédant le même vocabulaire, situées l'une à Paris, l'autre à 560 km de Paris. Le message "Il fait beau à Paris" apportera une information pratiquement nulle ($H_A = 0$) à la première, et non nulle H_B à la seconde, si effectivement il fait beau à Paris. S'il ne fait pas beau, l'information H_A très peu probable pour A tendra vers l'infini alors que H_B gardera le même valeur. On peut être tenté, dans ce cas, de défendre la Théorie en remarquant que l'idée "il fait beau à Paris" n'était pas dans la partie des répertoires commune à A, B et à l'informateur. Mais si l'idée était commune, il n'était nul besoin de la transmettre .... La vérité est que la Théorie ne s'applique pas à la transmission d'idées, pour lesquelles on ne peut pas faire d'analyse statistique préalable; à fortiori l'intersection des répertoires individuels ne peut être définie.

b) fonction du message lui-même

Supposons que le message suivant, constitué de 20 symboles binaires équiprobables, soit reçu d'une part par un système de décodage, d'autre part par un homme

11111111110000000000

Le décodeur physique devra à chaque instant choisir entre 2 symboles équiprobables. L'information est constante du début à la fin du message, et vaut un bit par caractère.

Le décodeur humain, par contre, remarque très vite la répétition du premier symbole. Dès le quatrième ou cinquième caractère le symbole 1 devient pour lui

...../

plus probable, donc apporte moins d'information. L'information devient ensuite presque nulle : $H = \varepsilon$ ($\varepsilon \rightarrow 0$), si le sujet ne sait pas qu'il y aura autant de 0 que de 1. Au onzième caractère, l'information remonte à $(1 - \varepsilon)$, puis décroît plus vite encore qu'au début. Ce comportement-type est évidemment sujet à autant de variations que d'individus : la probabilité du symbole suivant est établie d'après les symboles passés - en fonction de règles complexes propres à chaque individu - Le message peut être plus ou moins structuré, ou les structures plus ou moins évidentes. Par exemple, les deux messages suivants paraissent à première vue aléatoires :

0100100111010110101100101101

1101001101001101001101001 10100

Mais seul le premier ne possède pas de structure simple; le second est formé par la répétition du motif 110100 - Si l'individu récepteur s'en aperçoit, il recevra moins d'information du second message que du premier. Le rythme apparaît ainsi comme la plus simple des structurations. Mais, même dans un message binaire, d'autres types de structuration peuvent être effectués par le cerveau. Par exemple dans le message suivant

10110011100011110000111110000 0 etc.

on perçoit très rapidement l'existence d'une règle de composition. A partir de ce moment, la seule information que puisse nous apporter le message serait un changement dans sa structure : l'information perçue est liée au changement, à l'originalité, à l'imprévisibilité.

De tels phénomènes sont encore plus évidents avec des messages formés de lettres ou de phonèmes. Soient par exemple les messages

GAMGAMGAMPAMGAMGRAMGAMGOMGANNGAM

AIBOSOLEILGRSOLEILWSOLEILURPSOLEITLOU

Certains mots (GAM, puis SOLEIL) sont perçus d'emblée malgré une périodicité très floue et des erreurs. Pour expliquer, puis imiter ces propriétés fondamentales de la perception, il nous faudra recourir à un autre fil directeur, qui nous sera fourni par la théorie de la Forme.

II.5.6. Limites de la Théorie de l'Information

Avec les contraintes sémantiques de la communication parlée, nous découvrons les limites de la théorie. Il n'est plus possible de définir un répertoire des symboles ni une table de leurs probabilités. La notion même de probabilité, au sens mathématique du mot, doit être abandonnée au profit de la prévisibilité, quantité définie à chaque instant en fonction de l'individu récepteur, du message passé et du contexte, mais quantité trop complexe pour être mesurée, sauf dans quelques cas simples.

L'information, dans la communication entre individus, doit être comprise comme un taux d'imprévisibilité ou d'originalité; ce n'est plus une quantité mesurable, mais un concept, utile pour clarifier certains processus perceptifs. Pour éviter les querelles de mots, nous réserverons par la suite le mot "information" à la communication entre systèmes dont le code, le répertoire et le canal sont bien définis; dans

...../

le domaine perceptif, nous utiliserons plutôt les termes d'"imprévisibilité", d'"information perçue", ou d'"originalité".

Nous devrons, dans le chapitre suivant, examiner de plus près l'aptitude de notre esprit à structurer les messages, à détecter les périodicités, à comparer entre eux les éléments sémantiques. Ces propriétés seront étudiées dans le cadre de la perception auditive, et trouveront un début de cohérence avec la notion de Forme.

III - AUDITION ET PERCEPTION AUDITIVE

Nous examinerons dans ce chapitre les seuls aspects de l'audition qui nous semblent essentiels pour l'analyse de la parole. L'anatomie et la physiologie de l'audition, très largement décrites dans la littérature spécialisée, ne seront mentionnées que pour mémoire; mais nous insisterons sur les hypothèses d'adaptativité de l'oreille moyenne et de codage cochléaire. Nous verrons ensuite les particularités tenant à la perception des trois dimensions acoustiques : l'intensité, la fréquence et le temps. Nous serons conduits à exposer les grandes lignes de la théorie de la Forme, en lui associant la notion de catégorie perceptuelle. Enfin, nous décrirons un modèle de l'audition qui restera sous-jacent dans les études de synthèse et de reconnaissance automatique de la parole.

III.1 ANATOMIE ET PHYSIOLOGIE DE L'AUDITION

III.1.1. Description d'ensemble (bib. III-1, III-2, II-6)

La vibration acoustique recueillie par l'oreille externe subit diverses transformations avant d'être acheminée sous forme d'influx nerveux vers les centres supérieurs (fig. III-1 et III-2). Le tympan transforme la vibration aérienne en vibration solide; celle-ci est acheminée vers la cochlée au moyen de trois osselets - marteau, enclume et étrier - articulés d'une manière complexe, sur laquelle nous reviendrons. En cours de route, la vibration a changé de nature : le déplacement du tympan, en phase avec la pression acoustique, est devenu déplacement de la fenêtre ovale et du liquide cochléaire.

La cochlée (longueur 30 à 32 mm, section 1 à 3 mm2) est séparée en trois parties par la membrane de REISSNER et la membrane basilaire. La rampe vestibulaire et la rampe tympanique, remplies d'un même liquide (périlymphe), sont en communication à l'extrémité de la cochlée (apex) par l'intermédiaire d'un orifice de faible section (hélicotrème, environ 0,25 mm^2). La rampe vestibulaire communique également avec les canaux semi-circulaires, qui semblent jouer un rôle essentiel dans l'équilibration. Le canal cochléaire, situé entre les deux rampes, est rempli d'un second liquide de propriétés physiques voisines, l'endolymphe.

Le volume de liquide déplacé par le mouvement de la fenêtre ovale est donc transféré dans la partie inférieure de la cochlée, dont l'expansion est permise par la fenêtre ronde. Si le déplacement est de faible amplitude, la périlymphe peut être considérée comme incompressible et les parois osseuses comme indéformables; la fenêtre ovale et la fenêtre ronde sont en opposition de phase. Notons au passage que l'hypothèse de rigidité ne peut être admise qu'en première approximation, car le son (dans la bande 500 - 5 000 Hz) est également transmis à la cochlée par voie osseuse, avec une efficacité moindre que par la voie aérienne (seuils inférieurs d'une quarantaine de décibels) mais non négligeable; ce qui ne serait pas concevable avec des parois rigoureusement rigides.

...../

Sur la membrane basilaire se trouve l'organe de CORTI, ensemble complexe comprenant des cellules munies chacune d'une vingtaine de cils ancrés sur la membrane tectoriale (fig. III-2). Il semble que ces cellules (cellules ciliées) soient sensibles au déplacement relatif des deux membranes; les cellules ciliées externes réagissent lorsque les cils sont courbés dans une direction centrifuge par rapport à l'axe de la cochlée, et les cellules ciliées internes réagissent à un déplacement longitudinal et dirigé vers l'apex.

Les quelques 25 000 à 30 000 cellules ciliées (2 500 à 5 000 cellules internes, et 20 000 à 25 000 cellules externes) sont connectées à un nombre équivalent de fibres nerveuses (neurones du 1er ordre). Seulement les connexions sont loin d'être univoques : chaque cellule ciliée interne est reliée à une ou deux fibres nerveuses, et chaque fibre nerveuse est connectée à une ou deux cellules. Quant aux cellules ciliées externes, chacune peut recevoir un grand nombre de terminaisons nerveuses, et chaque fibre nerveuse peut être en relation avec un grand nombre de cellules (d'après LORENTE de NO - 1932 - cité par STEVENS et DAVIS). Ces deux "grands nombres" non précisés dans la littérature, peuvent atteindre plusieurs dizaines voire plusieurs centaines : les terminaisons d'une même fibre sont observables sur 1/2 tour du limaçon soit sur une distance comprise entre $\frac{1}{10}$ et $\frac{1}{4}$ de la longueur totale de la cochlée. Certaines fibres desservent simultanément des cellules internes et externes. Notons ici l'analogie profonde entre l'organe de CORTI et la rétine, dont les terminaisons nerveuses sont elles aussi multi et interconnectées. On peut également remarquer qu'il existe dans les deux cas deux types de capteurs différents, cônes et bâtonnets d'une part, cellules ciliées internes et externes d'autre part.

Tous ces neurones, qui forment la partie principale du VIIIe nerf crânien, aboutissent à un premier relais constitué par les noyaux cochléaires dorsaux et ventraux. Dans ces centres se trouvent deux types de connexions synaptiques. D'une part, les neurones du premier ordre sont reliés aux neurones du deuxième ordre, qui continuent leur chemin vers les corps genouillés. D'autre part, ils sont reliés "en parallèle" à des neurones d'une type différent, à axone court, entièrement localisés dans les noyaux cochléaires; ces neurones, qui forment un tissu analogue à la matière grise du cortex, sont tous interconnectés de manière très complexe. Il faut préciser ici que, si la transmission de l'influx nerveux le long de l'axone est assez bien connue, la transmission synaptique est beaucoup plus obscure : celle-ci est unidirectionnelle, requiert un délai de 0,5 à 3 millisecondes, et peut donner lieu à des effets de sommation ou d'inhibition. Il est d'ailleurs incorrect de considérer la transmission nerveuse comme purement impulsionnelle ("loi du tout ou rien"), car les effets de facilitation ou d'inhibition font appel à des potentiels sous-liminaires.

Avant de s'ancrer dans les noyaux cochléaires, certaines fibres provenant de l'oreille droite vont s'adjoindre au faisceau nerveux issu de l'oreille gauche, et réciproquement. Après le départ hors des noyaux cochléaires, les faisceaux nerveux changent de côté, dans l'ensemble, à part quelques fibres qui restent homolatérales.

Pour atteindre le corps genouillé, chaque voie nerveuse doit encore traverser un relais synaptique; celui-ci, univoque, n'est pas localisé à un niveau précis. Cependant, au niveau des tubercules quadrijumeaux postérieurs, on trouve des synapses en plus grand nombre, ainsi qu'une dérivation d'une partie des fibres vers le côté opposé.

L'influx nerveux est ainsi transmis jusqu'au niveau du corps genouillé médian par les neurones du 3° ordre. Le corps genouillé médian n'est pas un simple relais il reçoit également des fibres nerveuses provenant d'autres organes sensoriels, ainsi que du cortex. Les neurones du 4° ordre relient le corps genouillé médian à une zone précise du cortex auditif, appelée zone focale, dans laquelle se font les liaisons avec les tissus corticaux de l'aire auditive. Il est à remarquer que l'aire auditive, con-

..../

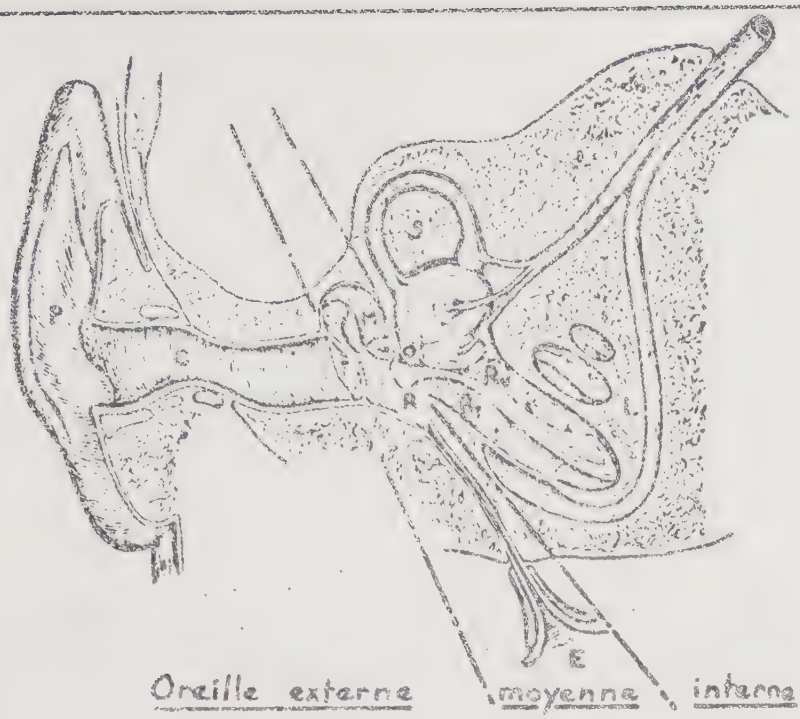
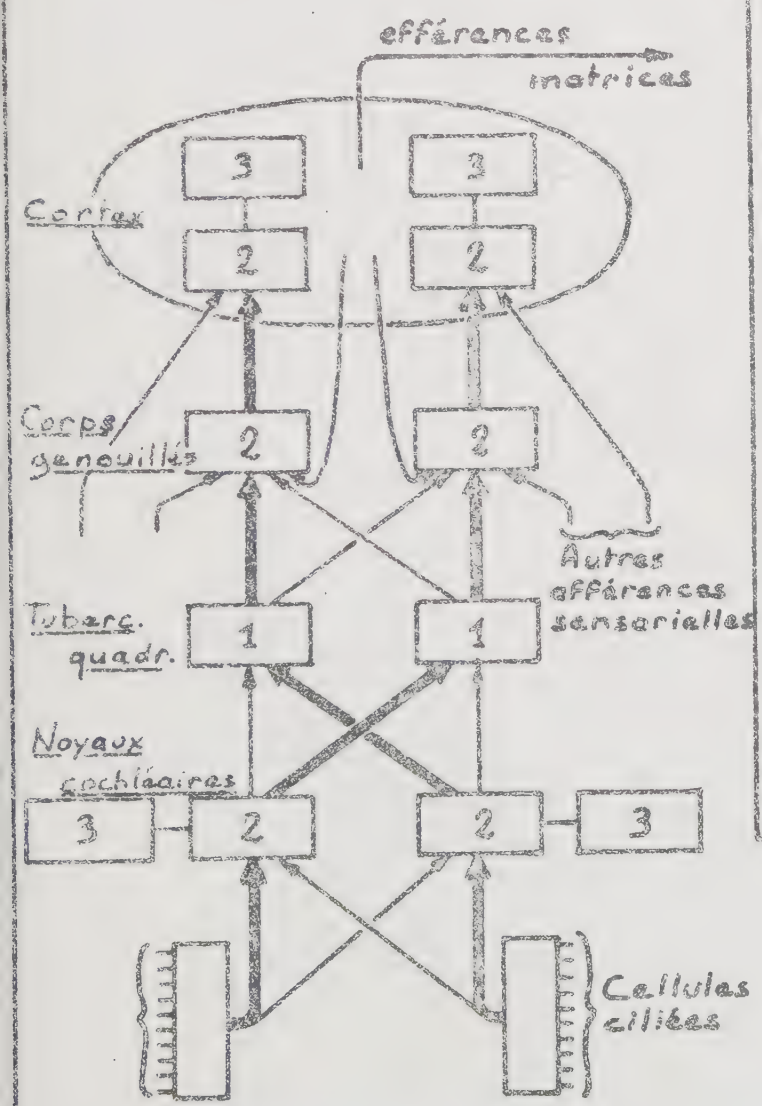


Fig III-1

Oreille (coupe d'ensemble). P pavillon ; C conduit externe; T tympan ; M chaîne ossiculaire ; O fenêtre ovale ; S canaux semicirculaires ; Rv rampe vestibulaire ; Rt rampe tympanique ; L limaçon ; E trompe d'Eustache; R fenêtre ronde.



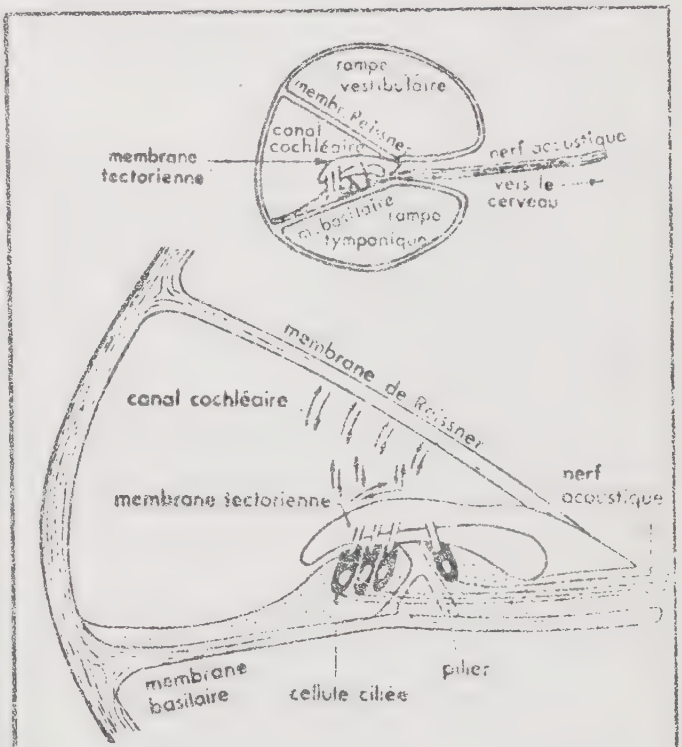


Fig III-2 Coupe de la cochlée (d'après E.LEIPP)

Fig III-3 Schéma des voies nerveuses supérieures.
1 : relais simples
2 : relais avec interconnexions
3 : tissus corticaux.

trairement à la zone focale, est une région extrêmement difficile à délimiter, tant par l'électrophysiologie que par la neurochirurgie, et qu'elle est le siège d'un grand nombre de connexions nerveuses non spécifiques à la fonction auditive.

La figure III-3 résume l'ensemble des considérations précédentes sous forme d'un schéma fonctionnel. Nous avons distingué, de manière quelque peu arbitraire, trois types de connexions neuroniques : les connexions univoques, ou simples relais (type 1), les connexions multiples, comme celles que l'on trouve dans la rétine et dans l'organe de CORTI (type 2), et les connexions faisant intervenir des neurones qui ne sont pas directement impliqués dans la transmission, comme celles des tissus corticaux (type 3) - Ce schéma met en évidence les nombreux croisements hétérolatéraux, qui font que l'on retrouve dans chaque aire auditive du cortex des informations provenant des deux cochlées.

III.1.2. Oreille moyenne et adaptation ossiculaire

On admet habituellement que le rôle de la chaîne ossiculaire est une adaptation d'impédance entre le milieu aérien de l'oreille externe et le milieu liquide de l'oreille interne. Si ce rôle n'est pas niable, vu les différences entre les propriétés élastiques des deux milieux, il faut bien entendre le terme dans un sens très large, car la transmission ossiculaire n'est pas plus linéaire en amplitude qu'en fréquence.

L'ensemble constitué par les osselets, le tympan, les membranes de la fenêtre ovale, les ligaments et les deux muscles tenseurs (du tympan et de l'étrier) est parfaitement dissymétrique (fig. III-4). Aussi un mouvement oscillatoire symétrique du tympan entraîne-t-il un mouvement dissymétrique au niveau de la fenêtre ovale (enregistrements optiques de DAHMANN, en 1929, rapportés par STEVENS et DAVIS). Cette dissymétrie, correspondant à une distorsion harmonique pour les phénomènes permanents, est d'autant plus marquée que l'amplitude du mouvement est grande, et que la liaison entre marteau et enclume est plus lâche. Or, la rigidité de cette liaison est évidemment fonction de l'état de tension du muscle tympanique (LEIPP, bib III-3). Il est donc permis de penser que la chaîne ossiculaire introduit une distorsion non linéaire en fréquence, le taux de cette distorsion étant variable selon l'amplitude du son et la tension du muscle tympanique.

Le mouvement de l'étrier change de nature selon l'amplitude de la vibration : aux fortes amplitudes l'étrier adopte un mouvement de pivotement autour du grand axe de sa platine, alors qu'aux faibles amplitudes il se comporte plutôt en piston. Ce phénomène, généralement interprété comme une protection naturelle contre les sons de forte amplitude, porterait sur une dizaine de décibels. En réalité, d'après LEIPP (bib. III-3) qui se base sur des considérations mécaniques et physiologiques, l'action combinée des muscles tenseurs du tympan et de l'étrier permettrait une accommodation de l'oreille dans une plage beaucoup plus importante, de l'ordre d'une trentaine de décibels. La contraction de ces deux muscles serait en partie volontaire et serait donc liée à la prévisibilité des signaux; cette hypothèse permettrait d'expliquer certaines observations pratiques relatives à la perception de l'intensité. Peut-être n'est-il pas indifférent de mentionner ici que le muscle de l'étrier est innervé par des fibres provenant du noyau du facial, en relation avec le noyau cochléaire via le complexe olivaire supérieur.

Une autre conséquence, mise en évidence par E. LEIPP, de l'action des muscles de l'oreille moyenne, est la déformation spectro-temporelle du signal : la raideur variable du système membranes-muscles-osselets entraîne un amortissement et une fréquence propre corrélativement variables : plus le système est raide, plus sa fré-

...../

quence propre augmente, plus son amortissement diminue. Il en résulte une possibilité d'adaptation de l'oreille non seulement au niveau sonore, mais aussi à la forme temps-fréquence du signal.

III.1.3. Le codage cochléaire

Le mouvement des fluides et des membranes de la cochlée est extrêmement complexe. Actuellement prévaut la théorie de BEKESY, qui, rejoignant l'idée des résonateurs cochléaires de HELMHOLTZ, suppose une véritable analyse fréquentielle dans la cochlée. Dans un modèle simplifié de la cochlée, la "camera acustica", BEKESY montre que le maximum d'amplitude vibratoire de la membrane se localise vers l'apex en basse fréquence, et vers la base en haute fréquence.

Ainsi la cochlée serait assimilable à un banc de filtres-résonateurs relativement amortis, de surtension constante $\frac{\Delta F}{F} \simeq 0,7$, échelonnés de manière logarithmique.

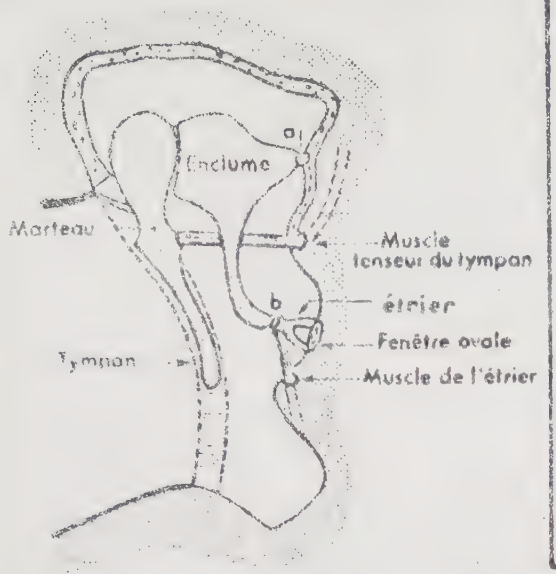
Diverses confirmations expérimentales sont venues étayer cette théorie. L'une des plus solidement établies est l'apparition d'un potentiel dit "microphonique" dans la cochlée, qui reproduit sensiblement la forme d'onde du son présenté au sujet. Par des études électrophysiologiques sur la cochlée du cobaye, STEVENS et DAVIS, en 1935, ainsi que CULLER en 1937, ont montré que l'origine de ce potentiel était localisée en un point de l'organe de CORTI évoluant de l'apex vers la base en fonction de la fréquence. Il est donc admis que le potentiel microphonique, dû essentiellement aux cellules ciliées, est l'image du déplacement des tissus supportés par la membrane basilaire.

A l'encontre de cette théorie nous pouvons, suivant LEIPP (bib. III-3, III-4) remarquer que la cochlée humaine est loin d'avoir la régularité géométrique du modèle de BEKESY (fig. III-5); que les expériences électrophysiologiques portent le plus souvent non sur des hommes, mais sur des cobayes ou d'autres animaux; que la cochlée étudiée est, soit perturbée par la pose d'électrodes, soit chirurgicalement endommagée; que l'on n'étudie que des phénomènes observables au microscope, alors que les phénomènes importants auditivement échappent à ce type d'investigation; que les sons utilisés sont généralement des sons sinusoïdaux, donc des artefacts du monde sonore habituel; que leur distorsion harmonique est dans une large mesure masquée par le moyen d'observation utilisé (oscilloscope); et que, en dernier ressort, une analyse fréquentielle au sens habituel du terme n'est pas absolument nécessaire pour expliquer la reconnaissance d'un signal acoustique.

Une localisation précise des fréquences sur la membrane basilaire - STEVENS et DAVIS vont jusqu'à mettre en parallèle les seuils différentiels de fréquence, de l'ordre de $\frac{\Delta F}{F} = 10^{-3}$, et la largeur, de l'ordre de 0,02 mm, occupée par une cellule ciliée interne - nous semble incompatible avec les faits expérimentaux suivant

- Le mouvement des liquides cochléaires et de la membrane basilaire est très amorti : tout son sinusoïdal met en vibration l'ensemble de la membrane, avec une très faible sélectivité. Ceci permet à notre oreille de saisir les phénomènes brefs, durant quelques millisecondes seulement, mais n'explique pas notre aptitude à discriminer des sons de fréquences voisines. WILKINSON et GRAY, puis BEKESY, supposent l'existence d'un système détectant le maximum d'amplitude vibratoire - mais peut-on détecter précisément le sommet d'une courbe très amortie; et cette supposition ne sort-elle pas du cadre expérimental fixé par les tenants de la théorie ? -

...../



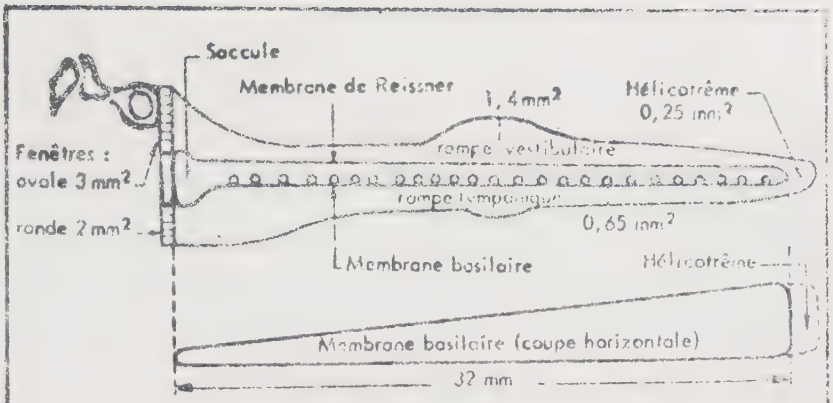


Fig III-5

Profil de la cochlée, déroulée (d'après E.LEIPP).

Fig III-4 Coupe de l'oreille moyenne (d'après E.LEIPP). Les ligaments ne sont pas représentés. Les muscles du tympan et de l'étrier permettent une accommodation en intensité et en timbre.

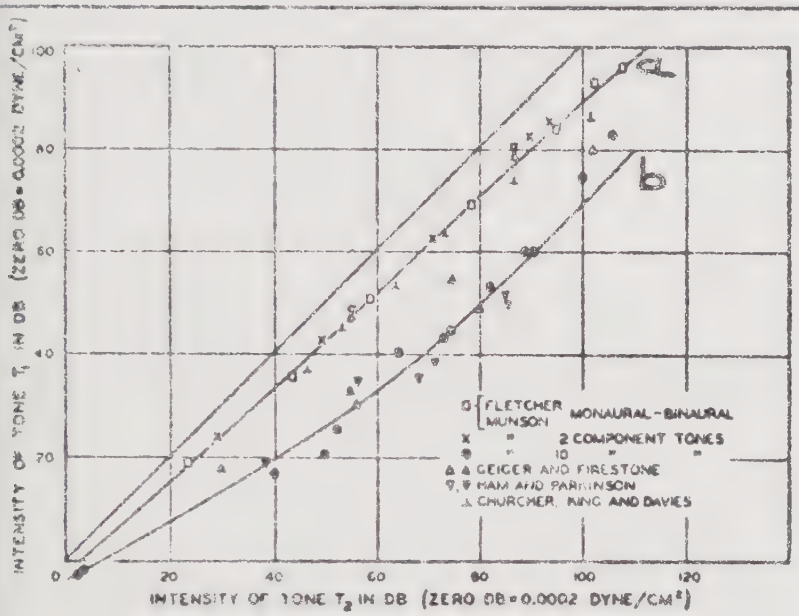


Fig III-6

Courbes d'intensité fractionnée (d'après STEVENS). En ordonnée est portée l'intensité d'un son T_1, qui paraît deux fois (courbe a) ou dix fois (courbe b) plus faible qu'un son de référence, dont l'intensité est portée en abscisse.

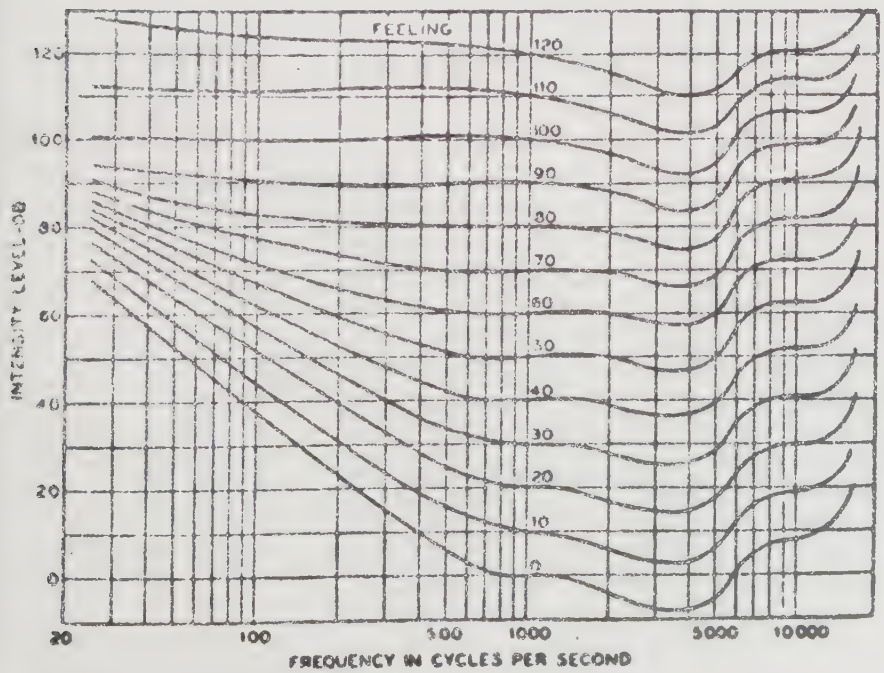
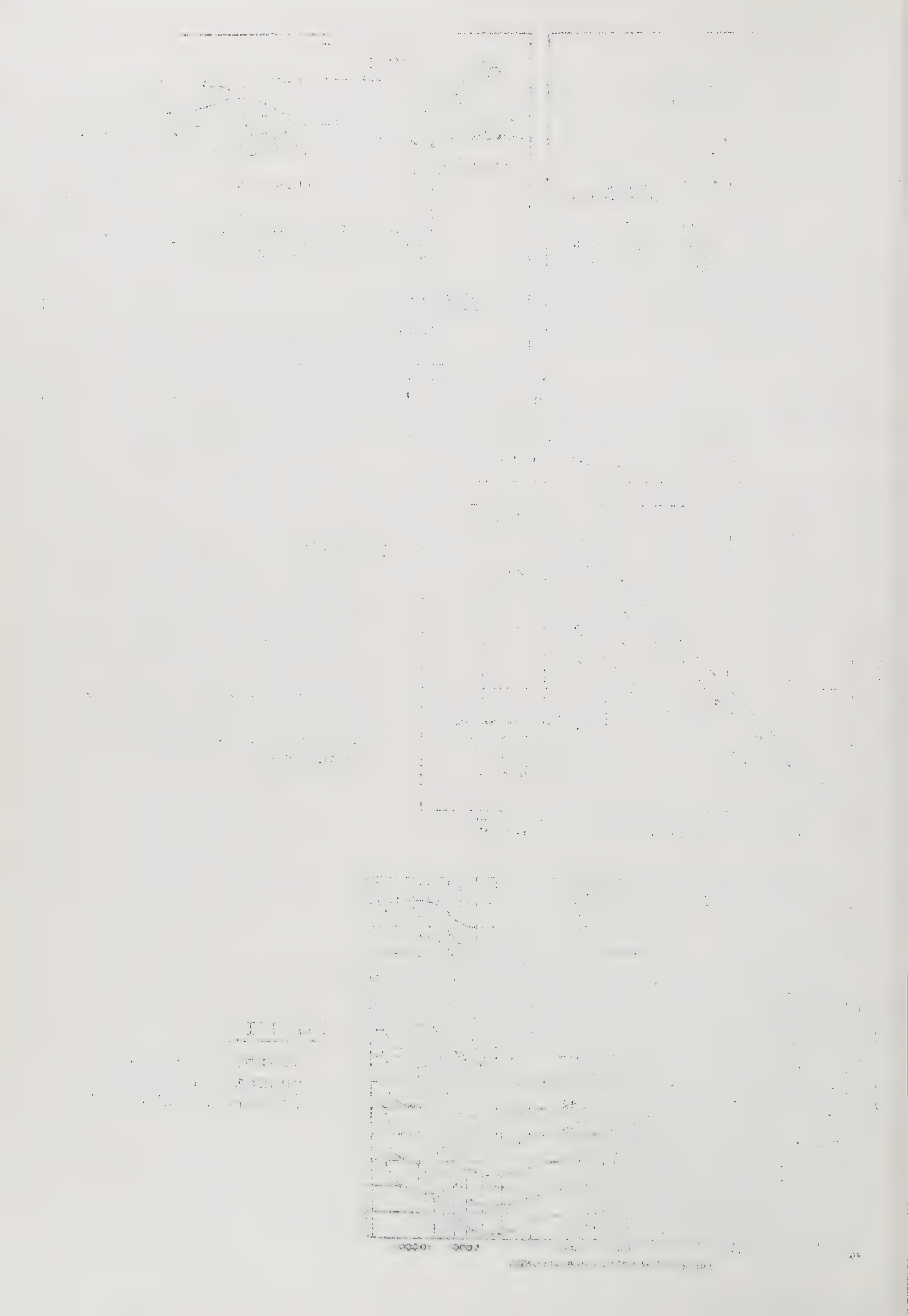


Fig III-7

Courbes d'égale sensation sonore (d'après FLETCHER)



- Nous avons évoqué plus haut l'extraordinaire réseau d'interconnexion des cellules ciliées et des terminaisons nerveuses, dans l'organe de CORTI. Il nous semble peu vraisemblable que des capteurs aussi étalés (jusqu'à 1/4 de la longueur de la cochlée) puissent transmettre sans ambiguïté une excitation précisément localisée : autant chercher à mesure des centimètres avec un millier de chaînes d'arpenteur !

Le potentiel microphonique n'est pas le phénomène important pour la transmission de l'information vers les centres supérieurs. L'influx nerveux envoyé dans un neurone du premier ordre provient du "potentiel d'action" délivré par une cellule ciliée au moment de son excitation, avec une latence de l'ordre de 1 ms. Les relations entre le potentiel d'action et le signal sonore sont beaucoup plus complexes que dans le cas du potentiel microphonique. Elles dépendent de toutes les caractéristiques acoustiques du signal (fréquence, intensité, durée) et de la cellule considérée.

On a cherché dans la succession des impulsions sur les fibres nerveuses du VIII° nerf un codage simple du message acoustique. On a pu constater ainsi qu'à un son de basse fréquence (inférieur à 500 Hz environ) correspond une suite d'impulsions synchrones sur certaines fibres. Mais la cadence de ces impulsions est limitée par la période réfractaire du nerf (environ 1 ms) ou des jonctions synaptiques (plusieurs millisecondes). Elle est également fonction de l'intensité (GALAMBOS et DAVIS, 1958), et perd son synchronisme dès que la fréquence dépasse 500 Hz. Si bien que la seule observation incontestable est un ensemble d'impulsions qui apparaît dans le nerf cochléaire quelques millisecondes après la présentation du signal acoustique.

Admettons que cet ensemble d'impulsions représente un certain codage du signal, plus exactement de l'objet sonore. Pour E. LEIPP le codage effectué par l'organe de CORTI n'a pas le caractère systématique d'une analyse fréquentielle, au sens métrique du mot. Il suffit qu'à un objet sonore donné corresponde une suite bien définie de configurations vibratoires de la cochlée, codées par les cellules ciliées. Peu importe alors la forme exacte du mouvement de la membrane basilaire ou la localisation d'un maximum, pourvu qu'un même objet sonore fournisse toujours à peu près le même ensemble d'impulsions, chez le même individu. Peu importent les différences anatomiques ou physiologiques observées d'un individu à l'autre, puisque le codage est individuel. Les problèmes posés par l'amortissement de la cochlée et par le grand nombre de cellules ciliées raccordées à une même fibre s'évanouissent : la membrane basilaire vibre dans son ensemble, son mouvement est codé dans son ensemble également. Ce point de vue ne contredit pas les résultats acquis jusqu'à présent : il n'est pas impossible que, pour les sons purs, les amplitudes maxima aient lieu près de l'apex en basse fréquence, près de la base en haute fréquence. Mais cette localisation est sans intérêt fonctionnel.

Le mouvement réel de la membrane basilaire est sans doute beaucoup plus complexe qu'il ne paraît : à en croire la théorie actuelle, on aurait uniquement affaire à des ondes longitudinales. Dans ce cas, les 20 000 cellules ciliées externes, sensibles au déplacement transversal de la membrane tectoriale, ne serviraient à rien ! ... La représentation classique de la cochlée développée selon sa longueur suggère une symétrie tout-à-fait artificielle : en réalité la section droite du canal cochléaire est loin d'une symétrie par rapport à un axe vertical (revoir fig. III-5), et il est tout-à-fait probable que les différences de pression statique (seules transmises au canal cochléaire) entre la rampe vestibulaire et la rampe tympanique entraînent des contraintes et des déformations transversales de l'organe de CORTI. De tels problèmes pourraient être approfondis au moyen d'un modèle mécanique réaliste de la cochlée.

...../

III.2. QUELQUES ASPECTS DE LA PERCEPTION AUDITIVE

III.2.1. Perception de l'intensité

a) Sensibilité différentielle et décibels

La constance expérimentale, dans de larges limites, du seuil différentiel d'intensité $\frac{\Delta I}{I}$ d'un son pur, produisant une même différence de sensation ΔS mène par intégration à la loi de WEBER-FECHNER

$$S = KL_n \left(\frac{I}{I_o}\right)$$

(I intensité physique
S sensation auditive
K constante
I_o intensité de référence

qui donne naissance au niveau N en décibels

$$N = 20 \log \frac{p}{p_o}$$

(H niveau sonore
p pression acoustique
p_o pression de référence (2.10^{-5} Pa)

Il semble que l'intérêt principal de la notation en décibels soit, non pas son origine différentielle, mais la commodité qu'elle procure dans l'estimation de la pression sonore : celle-ci peut évoluer entre le seuil de perception p_o et le seuil de destruction, quelque 10^6 fois plus. Cet intervalle peu maniable est réduit à 120 dB par la mesure logarithmique.

b) Paradoxe de l'intensité subjective fractionnée (fractional loudness)

Malgré la grande sensibilité différentielle de l'oreille (de l'ordre du décibel), qui lui permet de comparer avec précision deux sons de même nature et de niveaux voisins, les intervalles entre sons de niveaux très différents sont très difficilement appréciés. D'après la loi de WEBER-FECHNER un son de 80 dB devrait être apprécié comme deux fois plus "fort" que le même son, ramené à 40 dB. Il n'en est rien. De nombreuses expériences ont été faites dans ce sens (FLETCHER, STEVENS etc... fig. III-6), qui ont abouti à la notion de "sonie" et à un réseau de courbes normalisé reliant le niveau sonore objectif (sons purs) à la sensation d'intensité, mesurée en "sones". Cette notion s'est révélée parfaitement inutilisable en pratique. Pourtant, malgré le caractère indivisible des sensations, sur lequel nous reviendrons avec la théorie de la Forme, les courbes données fig. III-6 contiennent des renseignements intéressants; elles ont été établies en demandant à des sujets d'écouter deux sons purs sucessivement ou alternativement, et d'ajuster le niveau du second de manière à obtenir subjectivement une certaine fraction (ici 1/2 et 1/10) de la sensation produite par le premier. Dans l'ensemble, il suffit d'une réduction de l'ordre d'une dizaine de décibels pour produire une sensation moitié moindre, et d'une vingtaine de dB pour une sensation dix fois moindre. Donc, dans les conditions expérimentales, la perception de l'intensité suivrait plutôt une loi linéaire qu'une loi logarithmique.

...../

Les points portés sur la fig. III-6 sont eux-mêmes des moyennes d'observations individuelles, et proviennent d'auteurs différents. Nous avons toutes raisons de penser que, si ces points sont en bonne concordance, les relevés individuels présentent une dispersion considérable. Certains des auteurs (GEIGER et FIRESTONE, cités par STEVENS et DAVIS, bib. III-1) signalent que l'ordre de présentation des deux sons peut influencer les résultats. En tout état de cause, le principe même de ces expérimentations, qui tendent à chiffrer la sensation, peut être largement discuté.

c) Hypothèse de l'adaptation retardée

Pour expliquer l'observation précédente (perception non-logarithmique des différences d'intensité) nous allons faire l'hypothèse qu'il existe quelque part dans la fonction auditive un ou plusieurs mécanisme d'adaptation au niveau sonore, agissant avec un certain retard (de 1 ms à 10 secondes) - Plus le niveau moyen du signal passé et présent est élevé, plus la sensibilité vis-à-vis du signal à venir est réduite, et réciproquement. Ce système, qui sera simulé en reconnaissance de la parole (chapitre VI), a pour effet une modification complexe de la dynamique du signal : la différence d'intensité entre deux signaux quasi-stationnaires successifs est atténuée dans les parties permanentes, et augmentée pendant la transition. Sans doute encore simpliste, car l'adaptation présente également un aspect volontaire ou semi-réflexe, une telle hypothèse permet de concilier la sensibilité différentielle de l'oreille et les observations relatives au fractionnement de l'intensité subjective. Elle est d'ailleurs en relation avec l'idée de prévisibilité évoquée au chapitre II et avec ce que nous savons de la physiologie (sensibilité au changement).

Nous comprenons ainsi que le résultat d'une comparaison entre deux sons de même nature et de niveaux différents dépende non seulement du rapport de leurs niveaux en décibels, mais aussi

- de leurs niveaux absolus,
- de l'ordre de présentation et du taux de répétition,
- de leur transition, ou de la durée du silence qui les sépare,
- des caractéristiques du système régulé propre à chaque sujet : dynamique admissible et sensibilité, mémoire, rapidité, fatigue, etc...

d) Perception du niveau absolu

Cette question sera examinée plus loin, à propos de la perception catégorielle (§ III-3) - Nous remarquerons seulement ici que l'oreille ne se comporte pas comme un sonomètre, mais plutôt comme un comparateur : elle n'est pas plus capable de donner une mesure du niveau sonore que la main de donner la température précise d'un bain.

e) Perception de l'intensité en fonction de la fréquence

L'oreille présente un maximum de sensibilité aux alentours de 2 000 Hz : ce fait bien connu pourrait trouver son origine dans la résonance du conduit externe (courbe de WIENER et ROSS, montrant un gain d'une vingtaine de dB à 2 500 Hz) et dans la fréquence propre de l'oreille moyenne (1 400 Hz) mesurée par BEKESY - Cette dernière valeur est sans doute fortement variable selon l'état de la tension des muscles du tympan et de l'étrier (LEIPP,

...../

bib. III-2). Mais ces dispositions prises par la nature ne seraient pas très utiles si elles ne s'accompagnaient d'une plus grande résolution fréquentielle dans la même zone : les seuils différentiels d'intensité ΔI et $\frac{\Delta I}{I}$ (fig. III-8) sont en effet minima aux alentours de 2000 Hz et de l'ordre de 1 dB (RIESZ, FLETCHER et MUNSON).

Un autre aspect de la relation entre intensité perçue et fréquence est donné dans les célèbres courbes d'isosonie de FLETCHER et MUNSON (fig. III-7), qui montrent elles aussi un maximum de sensibilité dans la zone 500 - 5000 Hz.

III.2.2. Perception de la fréquence

a) Seuils différentiels et échelle logarithmique

Il est d'usage en acoustique de compter les fréquences selon une échelle logarithmique. Cette pratique se justifie par la constance approximative des seuils différentiels relatifs $\frac{\Delta F}{F}$, et par la perception des hauteurs musicales, qui admet un module d'octave.

Cependant la notation logarithmique présente de nombreux inconvénients, dont les principaux sont les suivants :

- elle masque l'équidistance réelle des harmoniques, à travers laquelle, selon E. LEIPP; l'oreille perçoit la hauteur du son (bib. III-5, I-2).
- elle minimise l'importance des fréquences aiguës par rapport aux fréquences graves. Dans la pratique acoustique - parole, musique ou bruit - l'essentiel des phénomènes perçus se situe presque toujours entre 1 000 et 4 000 Hz.

Un autre inconvénient - mineur - est que la courbe $\frac{dF}{F} = f(F)$ présente en fait un minimum assez marqué (fig. III-9), justement à 2 000 Hz - Si bien que son intégration ne mène pas exactement à une courbe logarithmique : la relation entre fréquence perçue ("pitch") et fréquence objective est donnée pour des sons sinusoïdaux par une courbe particulière (courbe des mels).

Ces considérations - et surtout la seconde - nous mènent à examiner les analyses de parole selon une échelle linéaire, en première approximation.

b) Zone sensible

Cette notion-clé se dégage de l'étude des courbes d'isosonie et de la répartition des seuils $\frac{\Delta I}{I} = f(F)$ et $\frac{\Delta F}{F} = f(F)$ - Dans cette zone du plan (F, I) centrée sur 2 500 Hz et 70 dB environ, on observe les valeurs minima des seuils différentiels, en fréquence comme en intensité. En d'autres termes, on peut définir des quanta bidimensionnels de perceptions (ΔI, ΔF), dont le nombre total est de l'ordre de 350 000 pour l'aire audible, et dont la densité est maximum dans la zone sensible.

Il ne faut pas se leurrer sur l'exactitude de telles spéculations : ces quanta ne sont pas indépendants, et ils proviennent d'études faites dans des

...../

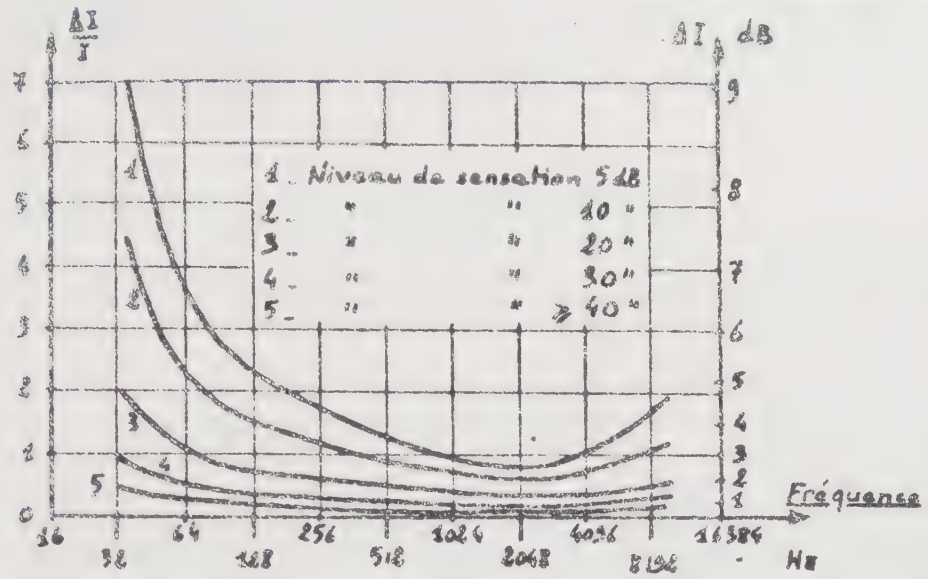


Fig III-8

Seuil différentiel relatif d'intensité, en fonction de la fréquence, pour divers niveaux de sensation sonore (d'après RIESZ, rapporté par STEVENS et DAVIS).

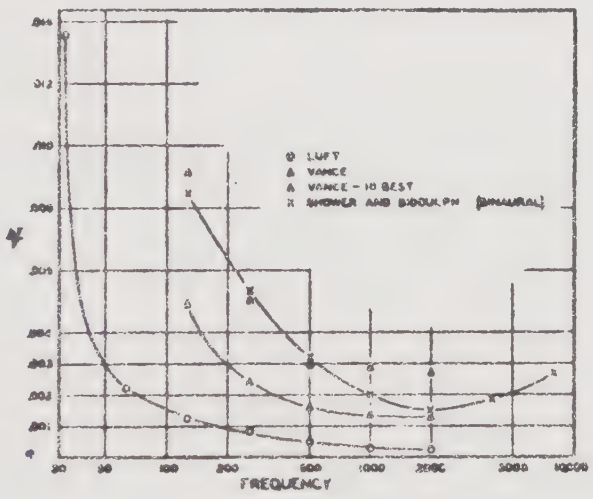
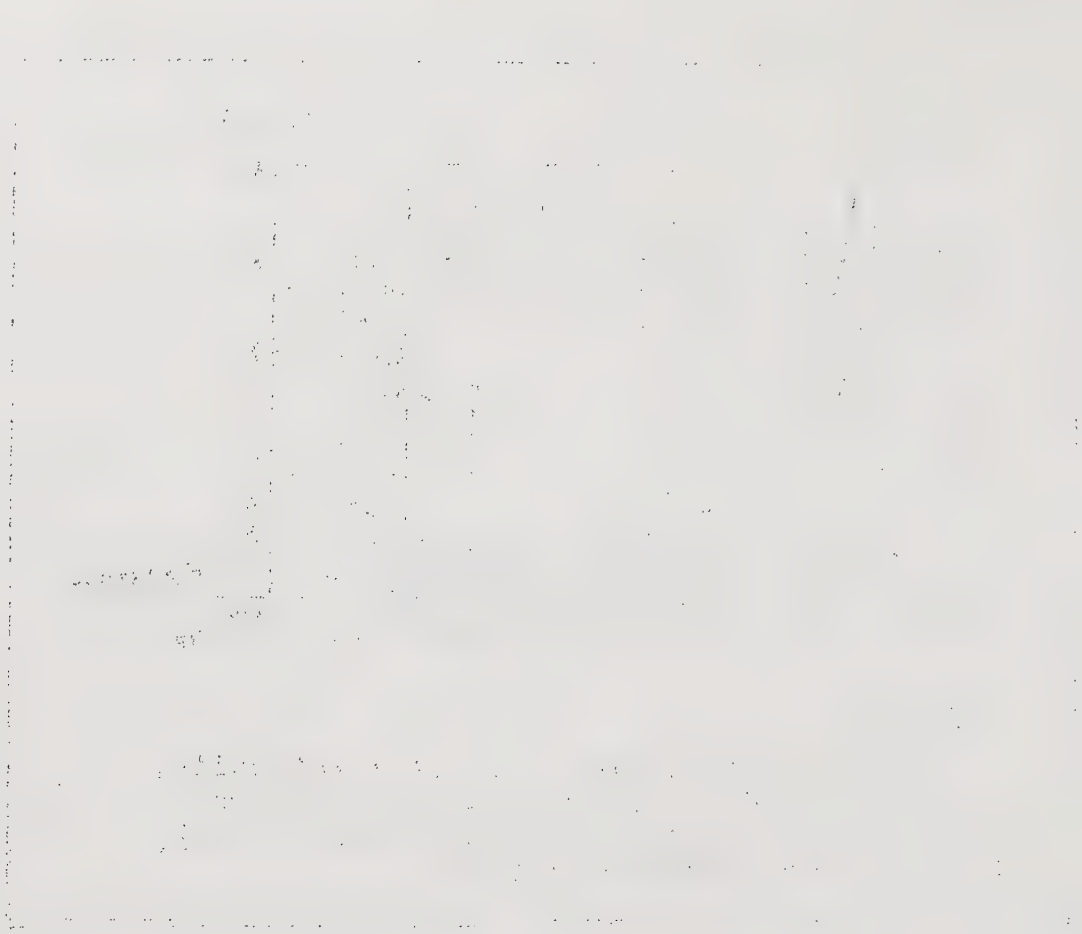
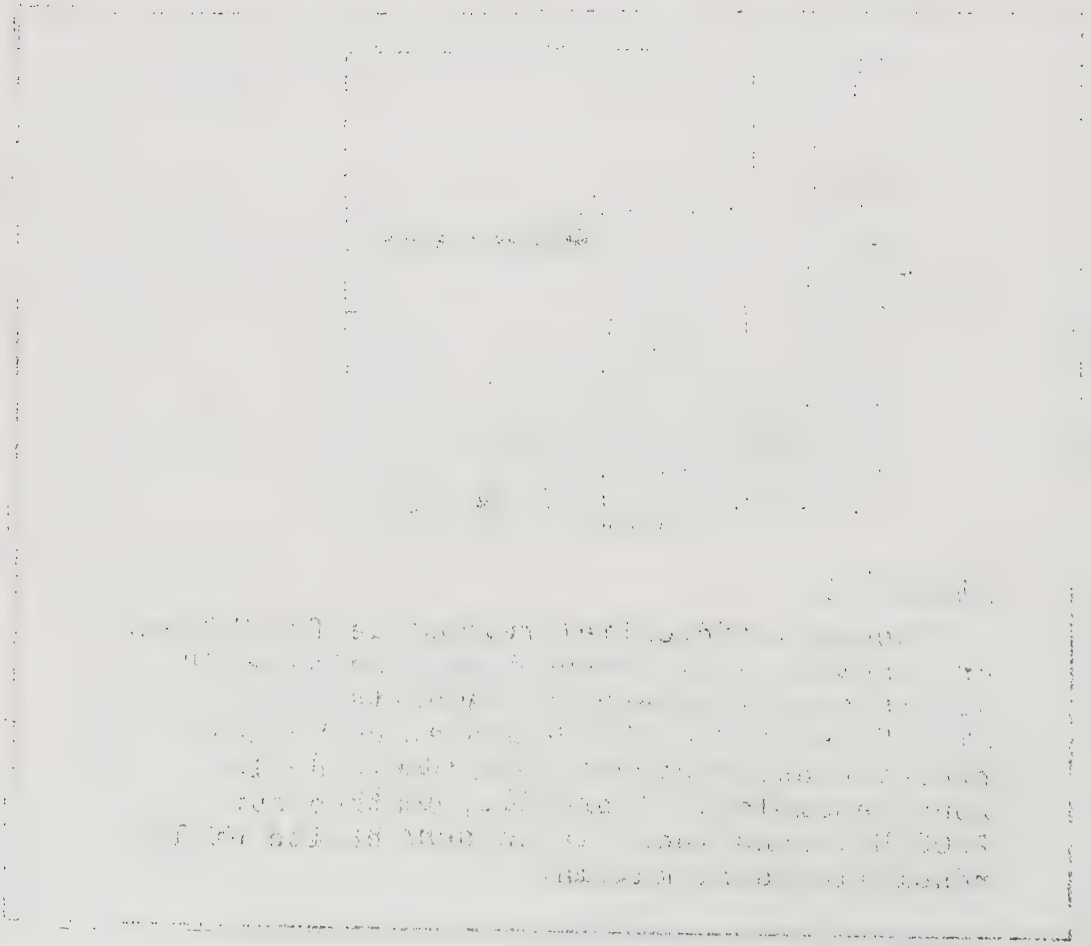


Fig III-9

Seuil différentiel relatif de fréquence, en fonction de la fréquence. Résultats de divers expérimentateurs, rapportés par STEVENS et DAVIS. Ces courbes, comme les précédentes, montrent l'existence de la zone sensible de l'oreille, centrée sur 2000 Hz, dans laquelle la sensibilité et la résolution sont maxima.





conditions artificielles (sons purs). La réalité est sans doute beaucoup plus grossière, comme nous le verrons au § III-3-4 - Mais cela n'affecte en rien l'importance de la notion de zone sensible, largement vérifiée dans la pratique.

c) Perception fréquentielle des sons complexes

Il semble que la sensation de hauteur prête à ambiguïté lorsqu'il s'agit de bruit, de sons harmoniques ou de sons non harmoniques.

On peut affecter une "hauteur" à un bruit de bande, plus ou moins floue selon la longueur de la bande et la forme de l'enveloppe spectrale. La hauteur d'un son harmonique, selon E. LEIPP, est perçue essentiellement à travers l'équidistance du réseau de raies. L'enveloppe spectrale est une composante importante du timbre. Cependant, si l'enveloppe possède un ou plusieurs maxima bien marqués ("formants") on peut leur attribuer une qualité de hauteur, surtout s'ils sont susceptibles d'évoluer en fréquence au cours du temps (guimbarde, parole, musique de formants). On voit ainsi qu'il n'y a pas de frontière nette entre hauteur et timbre; nous aurons l'occasion d'approfondir cette question à propos des anamorphoses de la parole.

Dans le cas des sons inharmoniques, la définition de la hauteur est encore plus problématique; il semble que les composantes les plus intenses ou les plus proches de la zone sensible soient prépondérantes (bib. III-5).

III.2.3. Perception du temps - Mémoire

Le temps est la dimension sonore la plus mal connue, parce que la moins étudiée. C'est cependant, à notre sens, la dimension la plus importante. Il faut d'abord s'entendre sur le terme : on ne perçoit pas le temps, grandeur abstraite, mais les événements qui le marquent; de plus, la perception du temps est soit incomplète (on ne perçoit que le passé d'un phénomène en cours), soit différée à l'aide de la mémoire; c'est pourquoi les caractéristiques de la mémoire déterminent en fait la perception du temps.

a) La constante de temps de l'oreille

Cette question donne lieu aux affirmations les plus variées : on trouve dans la littérature un éventail de valeurs allant de 1 à 200 millisecondes. C'est une querelle de mots. Qu'entend-on par constante de temps de l'oreille ? Les deux acceptions principales du terme sont les suivantes :

- Seuil différentiel de temps - C'est le plus petit intervalle séparant deux impulsions discernables l'une de l'autre. Il varie entre 1 et 10 ms selon l'intensité et les conditions expérimentales. La pratique de l'analyse et de la synthèse des sons nous montre à ce sujet que des modifications temporelles aussi petites que 4 à 5 ms sont parfaitement perçues, quoique non analysées.

- Constante de temps physiologique ou "épaisseur du présent".

  C'est la durée minima nécessaire pour qu'un son puisse être pleinement perçu, avec tous ses attributs, (hauteur, intensité, timbre, attaque, structure interne, etc...). On peut l'estimer à environ 60 ms. Cette

...../

valeur est à mettre en relation avec la plus basse fréquence perceptible (environ 16 Hz, période 66 ms), avec la persistance rétinienne (environ 55 millisecondes), avec la durée minima d'un phonème ou phonatome (de l'ordre de 50 ms), et avec le plus grand nombre de notes par seconde rencontré dans la pratique musicale - (12 à 15 notes par seconde, soit un intervalle de 67 à 83 ms).

b) Constante de temps, bande passante, et mémoire immédiate

L'ambiguïté de la question précédente provient de l'assimilation de l'oreille à un simple banc de filtres : comment concilier la sensibilité temporelle, de l'ordre de quelques millisecondes, avec l'aptitude à repérer très précisément la hauteur des sons, à quelques Hertz près (bib, III-6) ? Le produit $\Delta t \cdot \Delta F$ de cet analyseur serait de l'ordre de 0,01, ce qui est physiquement inconcevable (voir loi de KUPFMULLER, § II-1-2-4). Il y a là un faux dilemme, car on ne perçoit pas en même temps les plus fines nuances temporelles et spectrales; elles sont perçues en différé, grâce à la mémoire dite phosphorescente ou immédiate, alors même que le système oreille + cerveau continue à fonctionner. En d'autres termes, il semble que l'oreille soit capable d'appréhender un signal avec plusieurs fenêtres temporelles dont les durées s'échelonnent entre quelques millisecondes et quelques dixièmes de seconde (fig. III-10). L'oreille transmettrait ainsi aux centres supérieurs une description exhaustive du signal sonore, dont on peut donner une analogie visuelle en superposant deux sonagrammes de la même phrase (fig. II-16) obtenus l'un en bande étroite, l'autre en bande large. Cette hypothèse ne contredit en rien les données physiologiques : le mouvement relatif des membranes basilaire et tectoriale, image de la vibration acoustique, codé dans l'organe de CORTI, peut être transmis avec un retard variable selon les fibres, par le jeu des interconnexions synaptiques dans les noyaux cochléaires. On observe d'ailleurs, à ce niveau comme aux niveaux supérieurs, des "potentiels évoqués", ondes lentes durant plusieurs dixièmes de secondes, consécutives à une stimulation brève, qui indiquent pour le moins une certaine rémanence des stimulations acoustiques. Cette conception est assez voisine de celle de LICKLIDER qui, pour concilier l'analyse cochléaire en bande large (selon BEKESY) et la finesse de perception des hauteurs, admet que l'analyse cochléaire grossière est complétée par une analyse d'autocorrélation dans les fibres nerveuses. L'aptitude des circuits neuroniques à effectuer des opérations apparentées à l'autocorrélation est d'ailleurs assez couramment admise (bib. III-6).

Nous admettrons donc qu'un signal acoustique est "analysé" (au sens non-métrique, mais descriptif) selon plusieurs constantes de temps; cette analyse peut se réduire à un ensemble de comparaisons entre les états électriques des neurones "voisins", c'est-à-dire interconnectés. Une telle opération n'est pas sans analogies avec l'autocorrélation mathématique : rappelons que le spectre de puissance et la fonction d'autocorrélation d'un signal se correspondent dans une transformation de FOURIER. L'hypothèse selon laquelle les voies nerveuses sont susceptibles d'appliquer au signal des retards échelonnés s'accorde avec le mécanisme de régulation du niveau proposé au § III-2-1c et avec les observations relatives à la mémoire phosphorescente. Notons que ces retards et les interconnexions neuroniques multiples mènent à une interaction intime entre les aspects spatial et temporel de la vibration de la membrane basilaire, c'est-à-dire entre nos trois dimensions physiques de fréquence, intensité et temps, interaction largement démontrée par les études physiologiques. En conséquence, il faut s'attendre à ce que les traitements à effectuer dans ces trois dimensions en synthèse et en reconnaissance ne soient ni indépendants, ni fondamentalement différents; c'est ce que nous constaterons au sujet des anamorphoses, des comparaisons et des autocorrélations.

...../

c) Différents types de mémoire (bib. III-7, III-8, II-3),

On distingue généralement trois types de mémoires : mémoire phosphorescente, mémoire à court terme et mémoire permanente. La mémoire phosphorescente, dont nous avons déjà mentionné l'existence, permet une rétention limitée, qui se compte plutôt en éléments simples qu'en secondes. La rétention n'est pratiquement pas organisée, ni sélective; sa durée varie dans de larges proportions selon les individus et les expérimentations, et l'on peut seulement affirmer qu'elle est de l'ordre de quelques secondes, voire quelques dizaines de secondes. Par contre, sa capacité semble assez bien établie, puisqu'elle oscille entre 5 et 9 éléments du type chiffre, lettre ou mot, sans relation apparente entre eux. Remarquons néanmoins qu'il s'agit d'éléments symboliques, et que la capacité de 7 ± 2 traduit peut-être plus une limite de structuration du message qu'une limite de rétention proprement dite.

La mémoire à court terme semble être d'une autre nature. Elle permet de retenir les détails des événements pendant le temps où ils peuvent être encore utiles : en quittant ma table, je sais à quel endroit précis j'ai posé tel livre, tel dossier, tel objet - détails que j'aurai totalement oubliés demain ou dans une heure. Notons que la faculté d'oubli (ou de classement des souvenirs, ce qui revient presque au même) est complémentaire de la mémorisation : quels curieux infirmes serions-nous, si nous devions mémoriser sans discernement toutes les données sensorielles ! La première conséquence serait une très grande difficulté à retrouver un souvenir précis. La mémoire n'est pas comparable à un magnétophone, sauf peut-être pendant quelques secondes du passé immédiat.

L'existence des trois mémoires est une nécessité fonctionnelle. Mais il n'est guère possible d'assigner des limites précises à leur capacité, à leur durée d'action, à leur localisation. Il semble en tous cas que ces mémoires "actives" n'aient pas le même support chimique que les mémoires "passives" où sont stockées les données génétiques (J. MONOD, bib. III-9).

III.3 FORMES, STRUCTURES ET CATEGORIES

III.3.1. La théorie de la Forme (bib. III-10, III-11)

La théorie de la Forme (Gestalttheorie) est née en Allemagne à la fin du siècle dernier. Elle constitue une réaction contre la psychologie analytique du XIXème siècle, qui réduisait les faits de conscience et de conduite à des éléments, les sensations, et à leurs associations au sens de PAVLOV. Cette dernière tendance a été développée au XXème siècle par WATSON, dans la psychologie du Comportement ou "Behaviorisme" (bib. III-12). C'est de la psychologie du comportement que relèvent la majeure partie des études psychoacoustiques évoquées plus haut : selon cette doctrine, la psychologie doit établir des relations objectives entre des excitations sensorielles bien définies et mesurables (stimuli) et les réactions correspondantes de l'individu moyen (réponses), étudiées sur un groupe représentatif. Le comportement humain serait ainsi ramené à un ensemble d'automatismes simples dont la plupart seraient acquis par apprentissage depuis la naissance.

La psychologie de la Forme de VON EHRENFELS, KOHLER, WERTHEIMER, KOFFKA, etc.. se place sur un tout autre plan : elle considère les données sensorielles comme irréductibles à des éléments simples, sauf artefacts. Les faits psychiques s'organisent spontanément en Formes, ou totalités perceptives, dont les propriétés d'ensemble sont autre chose que la somme des propriétés des éléments constitutifs. Les Formes sont donc

...../

des données premières.

Voulant asseoir la théorie sur des bases scientifiques incontestables, les Gestaltistes ont cherché des exemples de Formes dans la physique, la chimie et la biologie, c'est-à-dire des phénomènes dans lesquels l'addition des parties ne possédait pas les mêmes propriétés que le tout. La chimie fournit facilement une quantité d'exemples, car le simple mélange additif de deux corps n'a évidemment pas les mêmes propriétés que leur combinaison. Mais il faut attirer l'attention sur la faiblesse de tels arguments, qui prêtent à polémique et ne semblent pas nécessaires pour constater le bien-fondé de la théorie dans d'autres domaines. Pour notre part, nous nous bornerons à considérer les Formes comme des totalités perceptives, c'est-à-dire des relations d'ensemble appliquées par l'esprit sur les données du monde extérieur.

Les théoriciens de la Forme ont posé un certain nombre de "lois", dont nous allons brièvement examiner les principales.

III.3.2. Lois des formes

a) Non additivité des parties

C'est en fait la définition même d'une Forme, illustrée par E. LEIPP (fig. III-10, bib. III-13). Une partie dans un tout est autre chose que cette partie isolée ou dans un autre tout.

b) Transposabilité et anamorphose

Certaines propriétés d'une forme se conservent dans des changements (ou anamorphoses) qui affectent d'une manière comparable toutes ses parties. Nous laissons ici volontairement dans le vague le terme "comparable", sur lequel nous reviendrons à propos de la perception catégorielle. Le caractère de transposabilité avait été mis en évidence en 1890 par von EHRENFELS, précurseur des Gestaltistes : une mélodie se conserve dans la transposition, alors même que toutes les notes sont différentes. Mais la transposition affecte tous les éléments de manière comparable, toutes les fréquences étant multipliées par un même coefficient.

c) Emergence d'une Forme à partir d'éléments

Etudiant les groupements de points ou de taches élémentaires, les Gestaltistes ont mis en évidence le rôle de la ressemblance et de la proximité des éléments dans leur structuration perceptive. On trouve aisément des illustrations de ces propriétés dans le domaine auditif. Il est plus facile, par exemple, de fabriquer une mélodie à partir de notes "proches" ou séparées par de faibles intervalles qu'à partir de grands intervalles; même chose pour les durées, qui sont en général du même ordre de grandeur. Quant à l'influence de la ressemblance, il suffit pour la vérifier de faire jouer chaque note d'une mélodie par un instrument différent, ou de faire dire chaque mot d'une phrase par une voix différente : la forme, mélodie ou phrase, perd de son évidence pour un auditeur non prévenu.

d) Loi de la "bonne Forme"

WERTHEIMER a remarqué que toute Forme, à l'intérieur de ses propres contraintes, tend vers la plus grande régularité, la plus grande symétrie, la plus

...../

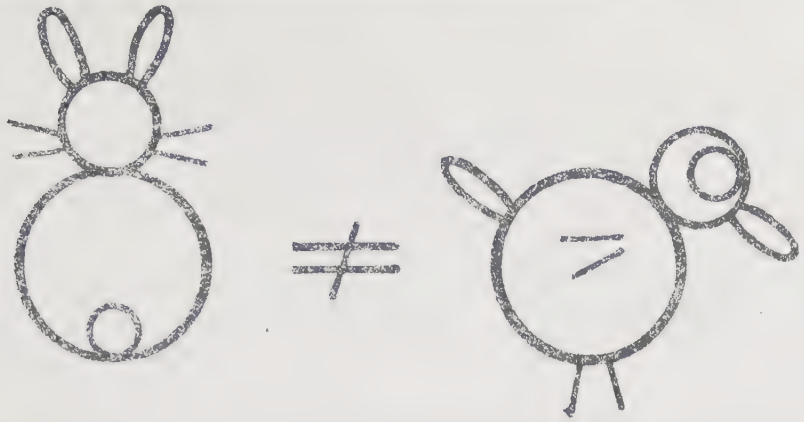


Fig III-10

Caractère de totalité d'une forme visuelle : les mêmes composants, assemblés différemment, donnent une autre forme.

Fig III-11

La figure de gauche est plus régulière, plus symétrique, plus continue ou plus compacte que les trois autres. C'est aussi la plus prégnante (loi de la "bonne forme").

Fig III-12

Figure ambiguë: on peut à volonté percevoir un motif blanc sur fond noir, ou vice-versa (d'après P.GUILLAUME).

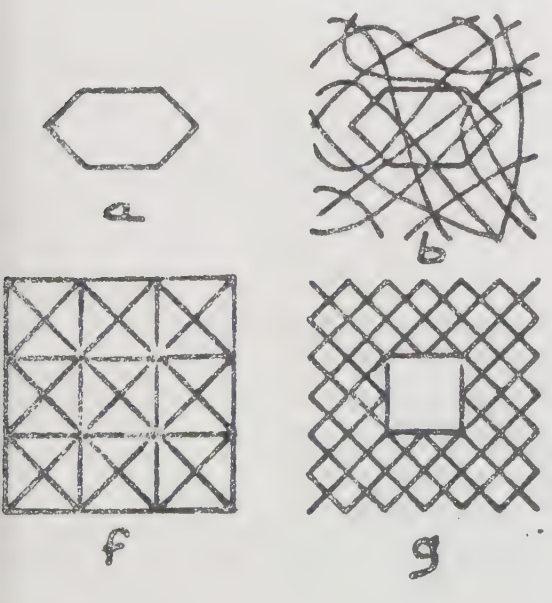


Fig III-13

Divers degrés d'émergence d'une figure sur un fond. a) Figure seule (fond clair) b),c) et d) Fonds laissant la figure inchangée. e) et f) Fonds intégrant partiellement ou complètement la figure. g) Apparition d'une nouvelle figure, plus prégnante : il est très difficile de distinguer la figure primitive (d'après P.GUILLAUME).

grande simplicité possibles. La figure III-11 propose diverses représentations du lapin de E. LEIPP : la forme la plus "prégnante" est celle qui présente les caractères ci-dessus.

Lorsque les conditions sont telles que l'esprit peut choisir entre différentes formes, il se tourne d'emblée vers la forme la plus prégnante, celle qui possède la plus grande cohésion intime.

Il nous faut remarquer ici que cette notion peut prêter à confusion : si le sujet s'attend, en fonction d'expériences antérieures ou de motivations internes, à percevoir une forme particulière, les conditions expérimentales sont faussées et le choix ne relève plus seulement de la prégnance, mais de la prévisibilité des formes possibles. C'est cependant le cas général de la perception réelle.

e) Formes et Superformes

L'assemblage de formes par juxtaposition ou superposition peut, si les critères de ressemblance et proximité sont suffisants, donner lieu à l'apparition d'une forme nouvelle ou superforme. En fait, les phénomènes perceptifs se prêtent généralement à l'établissement d'une hiérarchie : la phrase écrite est une superforme par rapport au mot, qui est lui-même une superforme par rapport au signe alphabétique, lequel se trouve être une forme composée de traits et de points. On pourrait aller plus loin dans les deux sens de cette hiérarchie, mais l'important est de montrer que la perception peut opérer à n'importe lequel de ces niveaux : le lecteur pressé lit "en diagonale", le lecteur attentif lit mot par mot, le jeune écolier déchiffre lettre par lettre ou syllabe par syllabe. Il en va de même pour le message parlé, saisi selon sa difficulté au niveau de la phrase, du mot ou de l'élément (cas de l'apprentissage d'une langue étrangère).

f) Formes fortes et formes faibles

Lorsqu'une forme s'articule selon plusieurs niveaux, tous les niveaux ne sont pas équivalents : une forme est dite "forte" lorsque ses parties (formes composantes) perdent leur individualité, et "faible" lorsque leur subordination au tout laisse subsister leur caractère propre. Une forme faible présente donc des régions de moindre cohésion, et des articulations au niveau des formes composantes.

La distinction précédente donne lieu à la même objection que la loi de prégnance : il est rare que toutes les formes possibles d'un phénomène soient également imprévisibles pour l'observateur. Une forme faible peut être détectée même si elle est intégrée dans une forme forte, si les observateurs savent qu'ils ont à rechercher cette forme.

g) Figure et fond

C'est là, avec l'idée de totalité, l'apport essentiel de la théorie de la Forme. Toute figure se détache sur un fond, qui possède des propriétés très différentes. La figure possède une organisation interne; le fond, même hétérogène, paraît informe et indéfini. La figure a des contours, qui en sont partie intégrante, alors que le fond n'en a pas. Les figures ambiguës, dans lesquelles figure et fond peuvent être perçus alternativement, permettent d'étudier ces propriétés. On vérifiera sur la fig. III-12 l'inversion figure-fond et le relief apparent de la figure par rapport au fond.

...../

L'émergence d'une forme sur un fond est un point central de la perception. La distinction d'une forme est d'autant plus difficile que son organisation est voisine de celle du fond. La forme a) (fig. III-13) est parfaitement distincte du fond en b), c), d), presque masquée en e) et totalement absorbée par le fond en f) - En g) elle est intégrée au fond, au bénéfice d'une autre forme de plus forte cohésion. Ces propriétés remarquables de la perception dans le domaine visuel sont exactement transposables dans le domaine auditif. E. LEIPP, dans une étude sur les signaux d'avertissement (bib. III-14), a montré que l'émergence d'un signal sur un fond pouvait être parfaitement mis en évidence avec la représentation temps-fréquence.

III.3.3. Théorie de la Forme et Structuralisme

La théorie de la Forme a prêté le flanc à de nombreuses critiques, dont la plupart concernaient des extrapolations philosophiques ou sociologiques, et n'ont pas leur place ici. Cependant, on a pu lui reprocher, avec juste raison, la donnée a priori de formes simples, le caractère statique, inné, des lois d'organisation. Et puis, il faut bien dire qu'elle compliquait singulièrement le travail des psychologues "scientifiques" ou behavioristes dont l'idéal était la définition et la mesure des ensembles stimulus-réponse.

a) Le Structuralisme

Le structuralisme a repris la notion de totalité différant de la somme des parties. Il a débaptisé la Forme, l'a appelée Structure, et lui a donné une portée plus générale en l'appuyant sur les mathématiques. La structure, au sens contemporain, est le lien qui unit les parties d'un tout; mais ce lien peut être aussi abstrait que l'on veut, et peut concerner des phénomènes disparates; une même structure servira par exemple à décrire un phénomène physique et un phénomène linguistique. Les mathématiques ont mis en évidence trois types de structures dites "structures mères", irréductibles entre elles, qui sont les structures algébriques (prototype: structure de groupe), les structures d'ordre (prototype : réseau) et les structures de voisinage ou topologiques. De plus, la structure est caractérisée par les idées de transformation et d'autoréglage (PIAGET, bib. III-15). Le problème de l'innéité des formes est alors contourné : une structure est définie par ses transformations et ses possibilités d'adaptation aux contraintes du phénomène considéré, plus que par son origine proprement dite, qui est mathématique, donc intemporelle.

Cette nouvelle perspective n'infirme pas ce qui a été exposé précédemment. Elle permet néanmoins d'introduire un facteur dynamique dans la perception, dont nous avons plusieurs fois déjà reconnu le besoin, au sujet de la prévisibilité des messages ou des formes.

b) Ordre proche et ordre lointain (bib. III-17, III-18)

On retrouve dans le structuralisme les notions gestaltistes de formes fortes, formes faibles, et superformes. Il existe donc des structures fortes, à grande cohésion, et des structures faibles, articulées en structures et superstructures. Mais, formulée différemment, cette distinction a abouti à la notion de distance d'ordre, tout à fait universelle. Considérons par exemple une voiture automobile : c'est une structure forte quant à la fonction, car chacune de ses parties est pratiquement indispensable au tout. Si nous en approchons, nous remarquons qu'elle est formée de pièces disparates dans leur forme, leur matière, leur destination : avec un certain "grossissement", la structure

...../

d'ensemble s'estompe, et l'on se trouve devant une structure faible. Mais en nous approchant un peu plus, nous découvrons que chaque pièce possède une certaine autonomie fonctionnelle et une certaine cohésion de forme et de matière. A ce niveau, nous retrouvons des structures fortes. Si l'on va plus avant, avec un microscope par exemple, on voit que cette cohésion disparait dans un enchevêtrement de fibres ou de grains. Mais ceux-ci, à leur tour, se révèlent ordonnés en molécules ou en atomes....

Ainsi un même phénomène peut présenter différents degrés d'ordre selon le "grossissement", c'est-à-dire selon les éléments considérés et leur distance moyenne. On pourrait décrire dans des termes semblables l'agencement d'un organisme vivant ou la structure du discours. Les expériences de parole interrompue (MILLER et LICKLIDER, bib. III-19) ont montré par exemple que l'intelligibilité de la parole était peu altérée par des coupures périodiques toutes les 50 millisecondes environ, alors que le même taux de coupure toutes les secondes ou toutes les 5 millisecondes affectait fortement l'intelligibilité. Ceci montre bien que le degré de cohésion du message parlé varie selon la distance d'ordre considérée.

Il convient de souligner la parenté entre la notion de distance d'ordre et le concept d'autocorrélation, pris dans son sens le plus large. Nous avons là deux désignations de la même propriété, qui est la ressemblance d'une structure avec elle-même, selon un certain décalage de l'un de ses paramètres. Nous en verrons plus loin une application, avec la courbe de stabilité temporelle du squelette sémantique. Mais ces notions s'appliquent également à d'autres domaines perceptifs, par exemple à la dimension fréquentielle de la parole : la perception de la hauteur tonale, c'est-à-dire du spectre de raies, relève d'un ordre proche, ou autocorrélation fréquentielle à court terme (quelques dizaines de Hertz), et la perception des formants et des bandes de bruit relève d'un ordre plus lointain (quelques centaines de Hertz). Si l'on suppose que le système nerveux central est apte à effectuer une autocorrélation temporelle du message, il est logique de faire une hypothèse comparable dans le domaine fréquentiel puisque les deux aspects sont présents (et sans doute indifférenciés) pendant la durée de la mémoire phosphorescente. Cette remarque n'est d'ailleurs qu'une formulation différente de l'hypothèse des retards échelonnés, proposée au § III-2-3-b.

III.3.4. La perception catégorielle

Nous avons souligné plus haut les difficultés rencontrées par la psychologie objective, par exemple au sujet de la perception de l'intensité : l'intervalle entre deux niveaux d'intensité ne peut être ramené à une somme de seuils différentiels. Si l'on demande à un auditeur une estimation de niveau absolu, il est incapable de donner une réponse précise en décibels; par contre, il peut donner une "impression", qui se traduit par l'un des jugements suivants : silence, son très faible, faible, moyen, assez fort, fort, très fort, vacarme infernal. Il serait illusoire de plaquer sur une telle échelle de valeurs une mesure physique objective, : les chiffres seraient susceptibles de variations considérables et non linéaires, selon l'oreille du sujet, ses conditions de vie, la nature des sons présents, les sons qu'il a entendus auparavant, son habitude de la formulation orale ou écrite, son état de fatigue, etc...

Il n'en reste pas moins que l'existence de telles catégories est universelle dans le domaine de la perception. MILLER est allé jusqu'à parler du nombre magique 7 ± 2, que l'on retrouve partout : dans la capacité de la mémoire phosphorescente, dans les jugements d'intensité (rappelons que l'échelle des intensités musicales com-

...../

porte les nuances de base pp, p, mf, f, ff), dans les jugements de durée (la notation musicale distingue 7 durées de base, de la quadruple croche à la ronde), dans les appréciations de qualité (p. ex. la docimologie usuelle utilise les qualificatifs : nul, médiocre, passable, moyen, assez bien, bien, très bien), et dans bien d'autres domaines (bib. III-6, III-20).

Pour les phénomènes qui présentent plusieurs dimensions ou aspects, POLLACK a montré que le nombre de catégories ne s'accroissait pas de manière multiplicative : un sujet capable de discerner 7 catégories d'intensité et 7 catégories de fréquence ne discernera pas 7 x 7 = 49 catégories de sons à deux dimensions (intensité et fréquence) mais une quinzaine seulement. Cela indique une forte dépendance entre la perception de l'intensité et de la fréquence, dépendance que nous avons constatée plusieurs fois déjà au cours de cet exposé, et qui se manifeste également sur la perception du temps. Nous sommes ramenés ainsi à la notion de totalité perceptive; la notion de catégorie est d'ailleurs implicite dans la théorie de la forme. En effet, la loi de la bonne forme de WERTHEIMER indique que toute forme est perçue d'abord comme la "bonne forme" la plus voisine. Les formes peuvent donc être ramenées à un certain nombre de formes-types, qui déterminent des catégories. Mais la théorie de la Forme n'a pas dénombré ces catégories, même dans le cas des formes géométriques, et n'a fait qu'une place minime à l'apprentissage de nouvelles formes.

III.4 MODELE FONCTIONNEL DE L'AUDITION

La généralisation actuelle du calcul numérique entraîne le développement des méthodes de simulation. La simulation d'un phénomène peut être effectuée au moyen d'un modèle concret (maquette électrique, mécanique ou autre), ou abstrait (programme, ou algorithme). Dans les deux cas, le but et la méthode d'investigation sont identiques : le modèle, bâti sur certaines hypothèses, donne des résultats qui, confrontés à l'expérience réelle, permettent d'améliorer les hypothèses. Cette méthode est une illustration du structuralisme : on admet que, lorsque les résultats sont satisfaisants, la structure donnée au modèle est fonctionnellement voisine de celle qui régit le phénomène étudié.

Le modèle, concret ou abstrait, est un outil de pensée. Il ne faut pas y chercher une correspondance anatomique ou physiologique précise; seules les opérations manifestement impossibles dans la réalité doivent en être proscrites. Si, par exemple, un modèle de l'audition incluait une mémorisation intégrale et permanente de l'information acoustique, ou un traitement instantané non causal, c'est-à-dire portant sur le futur du signal, ce modèle serait inadéquat.

Un modèle peut dans certains cas définir un phénomène aussi complètement et aussi rigoureusement qu'une équation. Mais dans le domaine de la perception, les modèles ne peuvent être encore que sommaires ou incomplets car, d'une part, les données expérimentales directes font pratiquement défaut (on ne sait pas, par exemple, sous quelle forme est stockée l'information) et, d'autre part, les données indirectes sont multiples et quelquefois contradictoires. Cependant, un modèle perceptif, même grossier, présente un triple avantage :

- il oblige à expliciter les hypothèses de travail,
- il donne un fil conducteur à l'étude,
- il permet de déterminer par ajustements successifs la structure d'ensemble du phénomène étudié.

...../

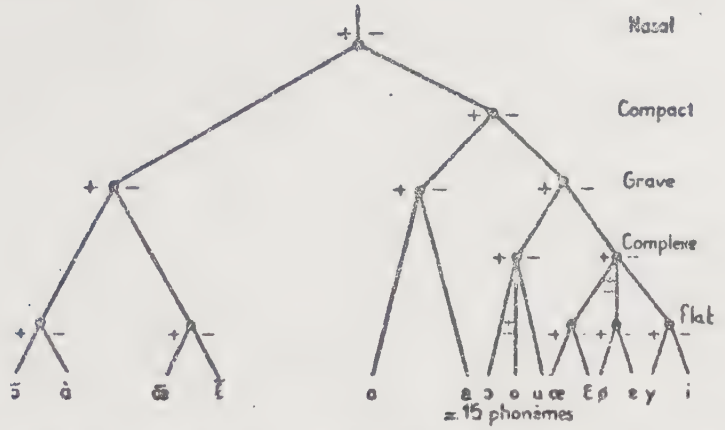


Fig IV-13 Traits distinctifs des voyelles françaises, avec exclusion du /ə/ , dit muet ou instable. Système maximum (d'après B.MALMBERG).

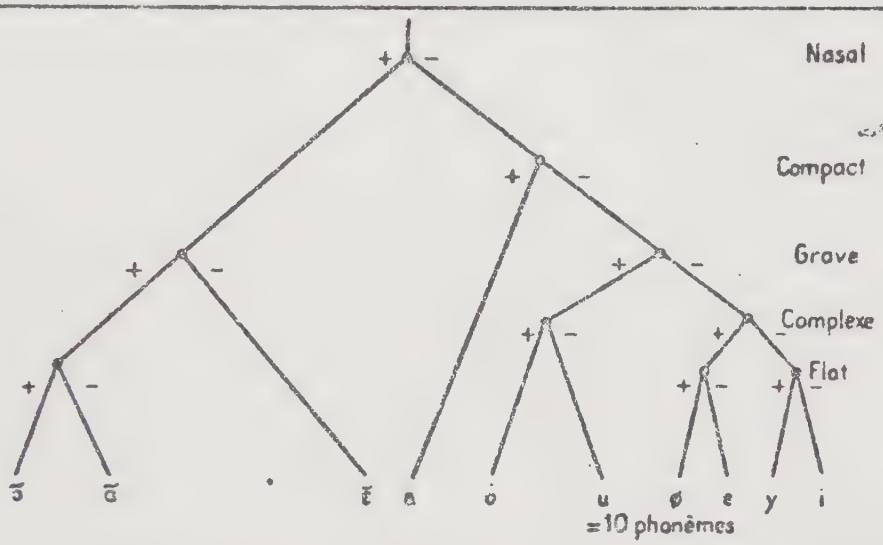


Fig IV-14 Traits distinctifs des voyelles françaises ; système minimum (d'après B.MALMBERG).

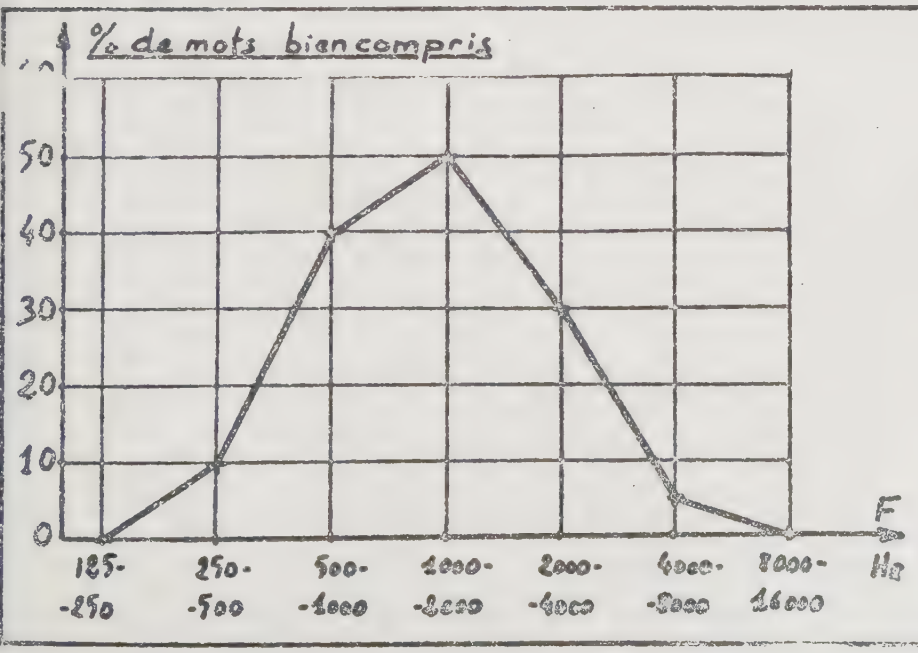


Fig IV-15

Intelligibilité de la parole, filtrée par bande d'octave. Moyenne établie sur 34 auditeurs. Listes de mots de FOURNIER, prononcés par deux locuteurs (masculin et féminin).

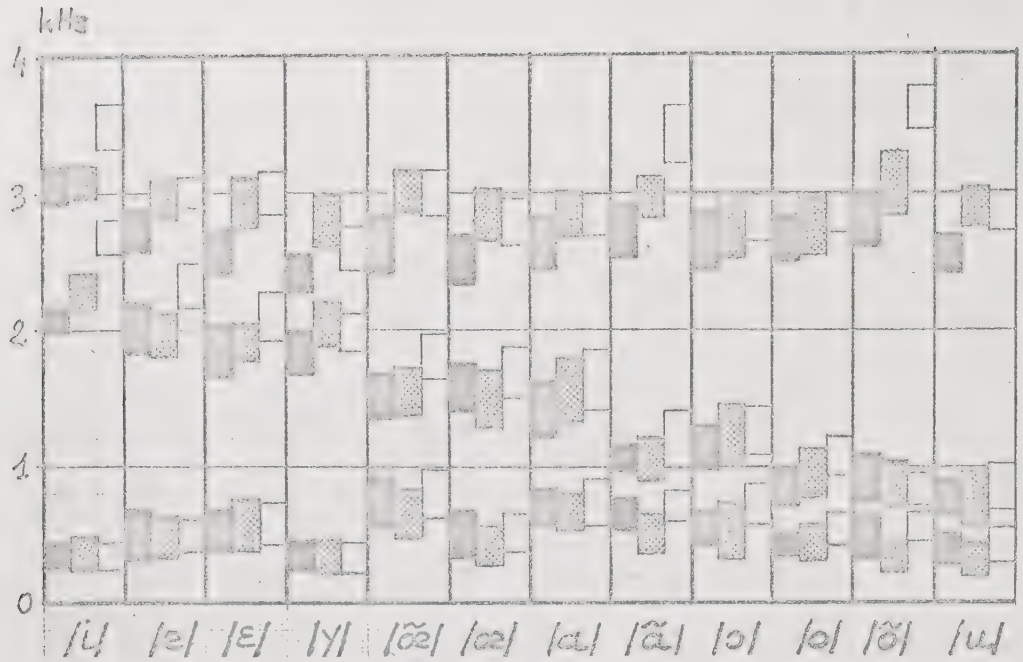


Fig IV-10

Fréquence des formants vocaliques, relevés en parole courante pour trois locuteurs : deux locuteurs masculins (en noir : JSL; en grisé : EL), et un locuteur féminin (MC, en blanc).

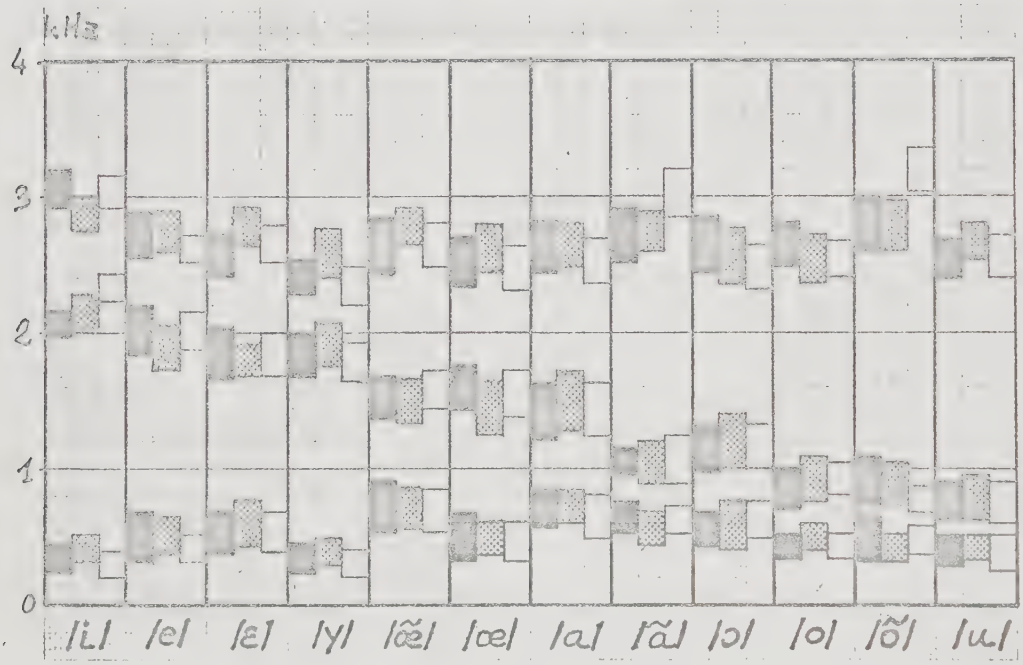


Fig IV-11

Mêmes relevés que précédemment, après application des anamorphoses $u_2(F)=1,15F-120$ $u_3(F)=1,15F$ pour les 2° et 3° locuteurs, par rapport au premier. Ces transformations réduisent la dispersion des relevés.

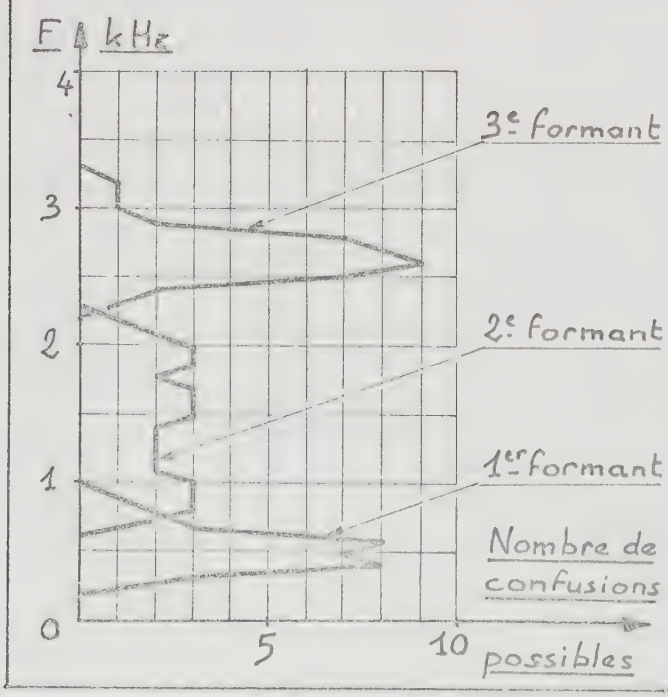


Fig IV-12

Etendue fréquentielle des trois premiers formants, établie à partir des relevés précédents (locuteur JSL). On a porté en abscisse le nombre de voyelles dont l'un des formants peut prendre une même valeur F. Cette figure montre que le second formant est le plus caractéristique.

matif, ayant pour sommets /i/, /u/, et /a/.

Le triangle des formants constitue une représentation élégante des voyelles, mais peu fructueuse. Il résulte en effet d'une démarche analytique, sans relation avec la perception. Il ne permet pas de comprendre comment des zones vocaliques peuvent s'interpénétrer, lorsque les résultats des mesures sont assortis de leurs zones de dispersion (fig. IV-9).

La forme allongée des zones de dispersion est par ailleurs fort édifiante. Lorsque le diagramme F_1 F_2 est représenté en échelles linéaires (cas de la fig. IV-9), on constate une orientation polaire des zones de dispersion. Dans l'ensemble, et à la précision du relevé près, on passe des points internes aux points externes de chaque zone en multipliant les fréquences F_1 et F_2 par un même coefficient. En d'autres termes, les aires de dispersion sont considérablement réduites si l'on tient compte d'un certain coefficient d'anamorphose entre les voix étudiées. Cette idée est confirmée par d'autres relevés relatifs à des voix masculines et féminines (voir par exemple TARNOCZI, bib. IV-4) : dans l'ensemble les points extrêmes se correspondent moyennant une multiplication des coordonnées par un coefficient de l'ordre de 1,6.

IV.2.2.4. Les formants des voyelles dans la parole courante

Nous avons cherché à savoir quelles étaient les fréquences des formants et leurs dispersions dans la parole courante. Pour cela, nous avons enregistré les mêmes phrases prononcées par trois locuteurs différents (deux voix masculines, une féminine) et comprenant au moins dix fois chacune des voyelles.

Sur la figure IV-10 sont représentées les zones dans lesquelles on peut trouver les fréquences centrales des trois premiers formants, relevées sur les sonagrammes, pour chaque voyelle et chaque locuteur. Les voyelles sont classées de l'aigu au grave, selon la fréquence du second formant.

Pour un même locuteur la dispersion des résultats est déjà importante : il n'est pas réaliste de définir la fréquence d'un formant à moins d'une centaine de Hertz près. La position des formants est d'ailleurs fonction de l'entourage phonétique : les variations des fréquences F_1 F_2 F_3 étaient parallèles pour les trois locuteurs.

D'un locuteur à l'autre, la dispersion devient considérable. Encore n'avons-nous pas là des voix très différentes : une voix d'enfant ou de basse donnerait des zones très éloignées de celles-ci.

La dispersion inter-locuteurs peut être notablement réduite si, prenant l'un des trois locuteurs comme référence (ici le locuteur n° 1), on pratique sur chacun des autres relevés une anamorphose particulière (fig. IV-11). La détermination de ces anamorphoses est effectuée en § IV.3.2.5., et mène aux transformations suivantes :

$u_2 (F) = 1{,}15\ F - 120$ pour le locuteur n° 2

$u_3 (F) = 1{,}15\ F$ pour le locuteur n° 3

A partir de ces relevés, nous avons également déterminé l'étendue fréquentielle de chaque formant. La figure IV-12 donne pour le locuteur n° 1 les ...../

mants. On distingue en français seize voyelles, auxquelles on attribue des positions articulatoires définies par les qualificatifs ouvert-fermé (il s'agit du degré d'ouverture de la bouche) ou antérieur-postérieur (il s'agit de la position de la langue). Par exemple, on oppose le /ɔ/ de "col" (/kol/) et le /o/ de "gaule" (/gol/) en disant que le premier son est ouvert, et le second fermé. Nous remarquerons simplement qu'il est parfaitement possible de prononcer ces deux mots avec une pipe dans la bouche, et de surcroit en immobilisant les lèvres, ce qui jette quelque suspicion sur la notion d'ouverture. Le tableau de la fig. IV-7 n'est donc donné que sous les plus expresses réserves, quant à la correspondance avec les positions articulatoires.

Il existe néanmoins certaines parentés sensibles entre ces voyelles, dont le tableau IV-7 rend partiellement compte. Nous dirons que /i/, /e/, /ɛ/ sont des voyelles aiguës, /y/, /ø/, / /, /œ/, /a/ et /ɑ/ des voyelles médianes, /u/, /o/, /ɔ/ des voyelles graves; à ces trois catégories sont rattachées les nasales /$\tilde{\varepsilon}$/, /$\tilde{œ}$/, /$\tilde{a}$/ et /$\tilde{o}$/, pour lesquelles le voile du palais est ouvert.

IV.2.2.2. Les fréquences des formants

Jusqu'à ces dernières années il était admis que les voyelles étaient perçues par l'intermédiaire des fréquences de leurs deux premiers formants. Encore faut-il préciser ce que l'on entend exactement par formant. Pour éviter toute ambiguïté, nous dirons qu'un formant est un pôle de la courbe de transfert du conduit vocal. En effet, si l'on se contente de définir le formant comme un maximum d'énergie dans le spectre, on risque de prendre en compte des maxima provenant du spectre d'émission des cordes vocales, et non du conduit vocal. Ces maxima parasites ne transportent pas d'information sémantique; ils évoluent selon le locuteur, selon l'intonation, selon le type de voix adoptée, et disparaissent en voix chuchotée. Ils existent généralement au niveau du premier et du second harmonique (fig. IV-8). On les détecte assez facilement sur les sonagrammes, car leur évolution est liée à l'intonation, c'est-à-dire au spectre de raies. Mais ils constituent une source d'erreurs considérable pour une machine chargée de détecter les formants, en particulier pour les voix féminines (cas 2 de la figure IV-8, par exemple); pour les éliminer en reconnaissance on pourra réduire fortement l'amplitude des fréquences graves (inférieures à 400 Hz environ).

Les fréquences des formants ont été mesurées par de nombreux expérimentateurs, essentiellement sur des voix masculines, pour lesquelles les formants sont mieux définis. A chaque voyelle correspond une combinaison particulière de ces fréquences. On a cru remarquer que chaque voyelle était caractérisée par les fréquences de ses deux premiers formants, ce qui n'est que partiellement exact. Cela est dû en partie au phénomène d'anamorphose que nous verrons plus loin : les mesures, faites le plus souvent à partir de sonagrammes, sont à précision $\pm \Delta f$ sensiblement constante, alors que l'anamorphose régulière (multiplication des fréquences par un facteur constant) se fait sentir surtout dans l'aigu. Il en résulte une plus grande dispersion sur la mesure des formants aigus qui, dans l'optique de fréquences formantiques parfaitement fixes, a incité les expérimentateurs à valoriser les deux premiers formants au détriment du troisième.

IV.2.2.3. Le triangle des formants

Il était dès lors tentant de représenter chaque voyelle par un point dans le plan F_1 , F_2 , F_1 et F_2 étant les fréquences des deux premiers formants. On constate ainsi que les points représentatifs se groupent dans un triangle approxi-

..../

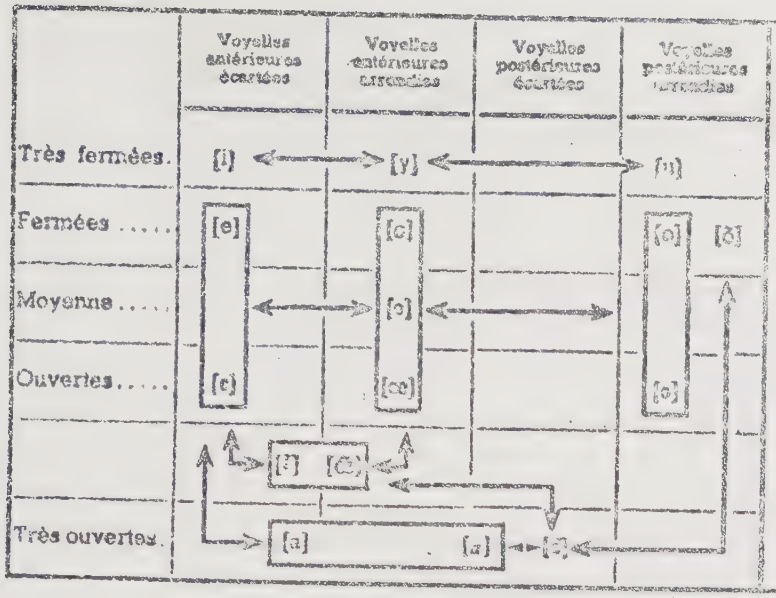
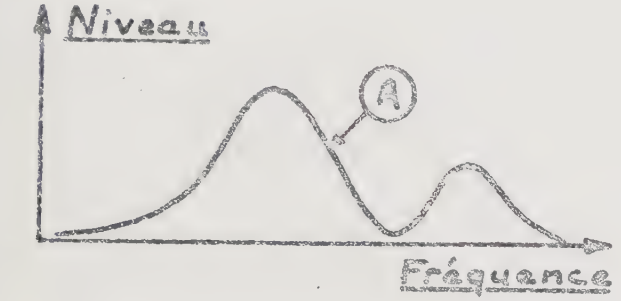


Fig IV-7

Système des oppositions vocaliques de la langue française (d'après P et M LEON).



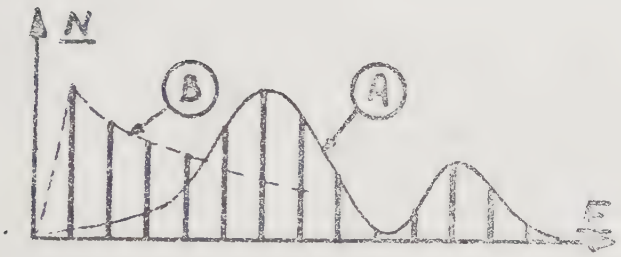


Fig IV-8

Vrais et faux formants. La courbe A est la courbe de transfert du conduit vocal ; ses maxima sont les formants. La courbe B est l'enveloppe du spectre propre des cordes vocales. Ses maxima sont des parasites pour le squelette sémantique. De haut en bas : voix chuchotée, voix voisée grave, et voix voisée aiguë.

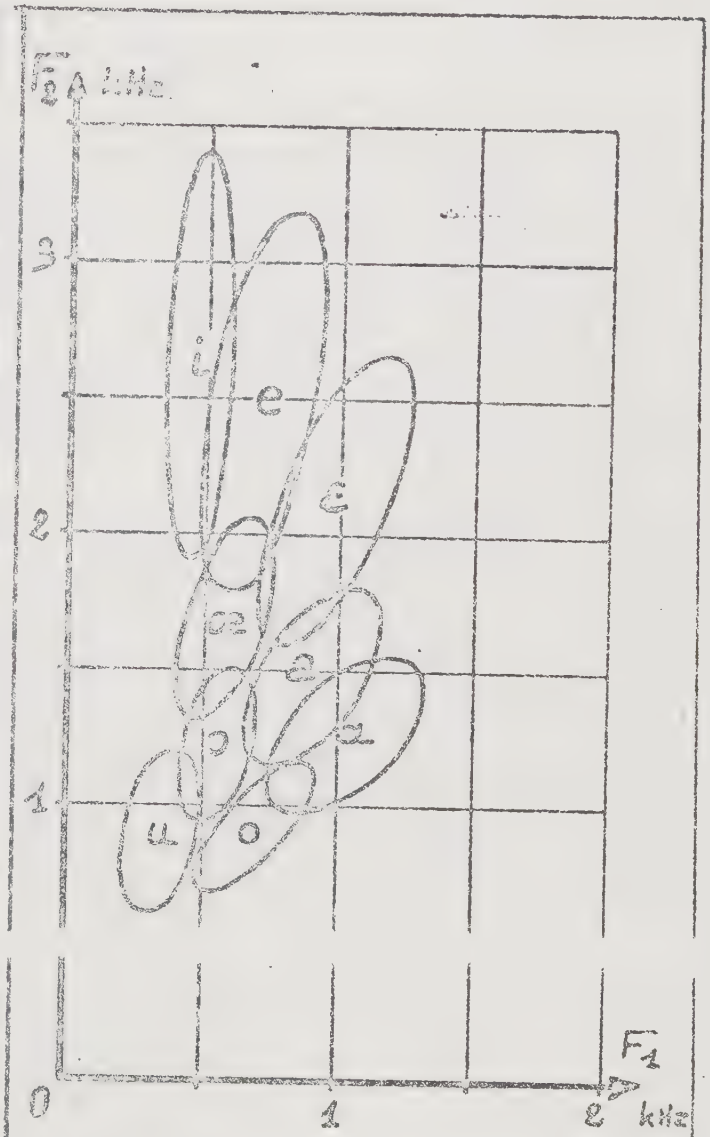


Fig IV-9

Les voyelles dans le plan F_1F_2 et leurs zones de dispersion (adapté d'après BARNEY et PETERSON).

dont la mécanique impose une dépendance entre notes successives. On peut même aller plus loin, et affirmer qu'un instrument est caractérisé plus par sa manière particulière de réaliser la forme d'ensemble repérée par la partition, que par son propre timbre.

Ce qui est vérifiable en musique l'est aussi en parole : les symboles phonétiques sont les paramètres abstraits du squelette sémantique, mais celui-ci peut être réalisé d'une infinité de manières selon chaque locuteur. Parmi ces réalisations, certaines font ressortir nettement les divers sons composants, comme le fait l'écriture imprimée, dans le domaine visuel, ou le jeu staccato en musique. D'autres versions du même message ne sont pas susceptibles d'une décomposition : seule subsiste la forme ou structure, qui constitue l'invariant sémantique. C'est le cas lorsque la parole est "mal articulée", lorsque l'écriture est manuscrite, lorsque le jeu est "legato". Mais le récepteur humain est capable de saisir la structure et éventuellement d'en donner la transcription paramétrée, c'est-à-dire la transcription phonétique ou alphabétique. C'est le problème de la reconnaissance automatique qui, à notre sens, ne doit pas se contenter de donner un résultat satisfaisant pour un locuteur sélectionné et habitué à sa machine, mais doit pouvoir s'adapter à des locuteurs et des contextes variés.

Dans ce sens, l'analogie entre les symboles phonétiques ("phonèmes") et les symboles alphabétiques ("lettres" du langage écrit) - est profonde. Cette sorte d'analogie est valable dans n'importe quelle langue, y compris les langues non alphabétiques. Dans les langues occidentales la notation écrite d'un message indique en même temps sa prononciation, donc est parente de sa transcription phonétique, ce qui constitue un grand avantage pour l'acquisition de mots nouveaux et la diffusion de la langue. Mais cette particularité ne doit pas nous induire en erreur : le phonème n'est pas une unité acoustique, sauf dans les cas de prononciation lente ou détachée, qui constituent des artefacts de la parole; par contre, le symbole phonétique est une unité de notation et peut-être de mémorisation du message parlé.

Il existe de nombreux systèmes de notation phonétique. Nous en utiliserons deux. Le premier est l'alphabet phonétique international, ou du moins le sous-ensemble de cet alphabet qui permet de noter la prononciation de la langue française. Le second est un système-maison, rendu nécessaire par l'utilisation du clavier alphanumérique usuel en informatique. Ce système, composé uniquement de lettres majuscules et de symboles spéciaux, a été établi en fonction d'une contrainte de programmation : chaque symbole phonétique devait être représenté par un seul caractère. Les deux systèmes ne présentent pas d'ambiguïté entre eux puisque le système international utilise des symboles minuscules, sauf peut-être en ce qui concerne la barre oblique / , séparateur dans le système international, et son "ON" (ballon) dans le système maison. Ce dernier utilise comme séparateurs les espacements ("blancs"), les virgules et les points. Les deux systèmes et leurs correspondances figurent à l'annexe I.

Il est maintenant nécessaire de présenter rapidement l'essentiel des notions phonétiques classiques, basées sur l'idée de phonème. Chemin faisant, nous découvrirons leur intérêt et leurs limites.

IV.2.2. Les voyelles

IV.2.2.1. Les voyelles françaises

Les voyelles sont généralement étudiées en régime permanent, à l'état isolé. Elles sont liées à l'existence dans le spectre de plusieurs résonances ou for-

...../

différence de sens provient uniquement de l'opposition entre /l/ et /r/. Puis on définit une opposition entre d'autres mots, par exemple "malin" /m a l $\tilde{\varepsilon}$ / et /m a t $\tilde{\varepsilon}$ /, pour montrer le rôle de /l/ et /t/, etc... Le phonème est donc défini comme le segment acoustique minimum susceptible de faire changer le sens d'un mot. Cependant, cette méthode repose sur des hypothèses par trop fragiles: d'une part elle se base sur la notation phonétique qu'elle prétend définir; d'autre part elle fait une place minime au contexte. Jamais un auditeur réel ne se trompera sur le sens des mots "lampe" et "rampe" même s'ils sont mal prononcés, sauf jeu de mots ou mauvaise connaissance de la langue. Et, dans les cas où une ambiguïté est à redouter, le locuteur emploie généralement un autre mot, une périphrase, ou communique l'information sémantique sous une autre forme, qui peut être non verbale. Le contexte n'est jamais absent d'une communication entre individus. Par contre, les machines à synthétiser et à reconnaître la parole ne sont pas encore aptes à utiliser le contexte; mais ce qui est une imperfection dans le domaine de la communication constitue un avantage considérable dans le domaine de la phonétique objective. Nous aurons l'occasion de revenir sur ce point.

IV.2.1.4. Phonème et position d'articulation

Nous savons par l'étude du conduit vocal que celui-ci fournit pour certaines positions un ensemble de résonances et anti-résonances qui sont les pôles et les zéros de sa courbe de transfert. Il est tentant d'associer un phonème, élément de parole, à une position-type du conduit vocal. Cette correspondance est possible, dans la limite des variations individuelles, pour les phénomènes quasi-stables que sont les voyelles et les consonnes fricatives. Aux phonèmes de nature transitoire, on associe la position articulatoire observable pendant la compression (pour les plosives),pendant la nasalisation (pour les nasales), etc...

Il faut souligner le caractère rudimentaire de ces correspondances, qui se réduisent à un ensemble de distinctions du type "antérieur, médian, postérieur". On pourra nous objecter que certains synthétiseurs fonctionnent avec des conduits vocaux analogiques; cela prouve qu'ils permettent une certaine synthèse, mais cela ne prouve pas qu'ils rendent compte du fonctionnement acoustique de l'appareil phonatoire. En d'autres termes, il nous semble utopique de concevoir un système de reconnaissance automatique basé sur un fonctionnement-type du conduit vocal, étant donnée la multitude de positions et de mouvements possibles au regard de la relative simplicité du squelette sémantique de la parole.

IV.2.1.5. Notation phonétique

L'intérêt principal de la notion de phonème réside dans la possibilité d'une notation phonétique, plutôt que dans la description acoustique ou phonologique de la parole. Une transcription phonétique soignée, comportant les principales indications de rythme et de mélodie, permet de préserver l'information sémantique et une partie de l'information esthétique du message. Mais, si l'on veut éviter des échecs en synthèse et en reconnaissance, il ne faut jamais perdre de vue le caractère abstrait des symboles phonétiques. La notation phonétique est comparable par certains aspects à la notation musicale : on chercherait en vain dans la réalité musicale telle note expressément indiquée par la partition; la note en question n'existe pas en tant que telle, mais constitue un paramètre de la "forme" mélodique dont elle fait partie. Ceci est particulièrement flagrant lorsque la phrase musicale est jouée legato, mais peut être constaté dans le jeu de la plupart des instruments (mis à part les instruments électroniques,bib. IV-3),

...../

l'occlusion.

La remarque précédente illustre un problème constant de la phonologie expérimentale : on n'observe commodément les phénomènes qu'en régime permanent; mais la parole est mouvement, et la réalité est différente. Nous verrons plus loin que cette difficulté est à l'origine de la dualité entre phonèmes et phonatomes.

IV.2 STRUCTURE PHONETIQUE DE LA PAROLE

IV.2.1. Quelques définitions et remarques

IV.2.1.1. La double articulation (bib. IV-2)

On retrouve en linguistique la distinction déjà signalée entre forme et superforme, ou structure et superstructure. Le message parlé est composé d'éléments, phonèmes ou phonatomes, d'une part,et d'autre part d'assemblages de ces éléments, que nous appellerons en première approximation des mots, et que les linguistes appellent, selon le cas, des monèmes, des sémantèmes, des morphèmes, des lexèmes ou des syntagmes nominaux.

La première articulation est donc l'articulation en mots, la seconde est l'articulation en éléments. La différence est comparable à celle qui existe en Informatique entre le "mot-machine" (mots de 16 bits, par exemple) et les symboles qui le constituent. L'intérêt est le même, et réside dans la possibilité de fabriquer un nombre quasi-infini de mots différents avec un nombre très faible d'éléments.

IV.2.1.2. Le point d'articulation

Nous avons utilisé cette notion plus haut, sans la nommer : pour les consonnes fricatives et plosives, le point d'articulation est assez bien défini, et correspond au rétrécissement du conduit vocal. Pour les autres consonnes et les voyelles, le point d'articulation désigne l'endroit où le profil de la langue est le plus proche de celui du palais. Cette notion, confuse du point de vue acoustique, a pour seul intérêt de repérer, très imparfaitement, la position de la langue; elle est universellement utilisée en phonétique, mais tend à faire croire

- que la forme du conduit vocal peut être décrite par un seul paramètre;
- qu'à un son donné correspond une position-type, valable quel que soit le locuteur.

Ces deux propositions sont évidemment très schématiques .

IV.2.1.3. Phonème et contexte

On définit souvent le phonème par la méthode des oppositions phonétiques : entre "lampe" / l $\tilde{a}$ p/ et "rampe" / r $\tilde{a}$ p/ , par exemple, on admet que la

...../

issue de la glotte, qui produit une pression acoustique homogène et en phase dans chaque section droite.

Une telle approche du problème conduit, sur des modèles mathématiques simplifiés (bib. II-6) à mettre en évidence plusieurs résonances pour une position donnée du conduit vocal, qui coïncident sensiblement avec les formants des voyelles, mais auxquelles on ne peut pas attribuer une localisation précise. Les calculs à effectuer sont passablement compliqués, même dans le cas de phénomènes permanents; ces dernières années ont vu se généraliser la notion de conduit vocal analogique (réseau électrique) commandé par calculateur en vue de la synthèse de la parole.

b) Génération de bruits dans le conduit vocal

Le larynx n'a pas le monopole de la production du son dans l'appareil phonatoire. Le conduit vocal est à l'origine des bruits de friction dans les sons /s/, /ʃ/, /f/ , /z/, /ʒ/, /v/, et des bruits d'explosion de /p/, /t/, /k/, /b/, /d/, /g/ - On peut même lui attribuer des bruits parasites comme les impacts et les clics accompagnant les mouvements de salivation; ces parasites, fréquents chez certains locuteurs, sont très sensibles lors de l'écoute de parole enregistrée à faible distance.

Les bruits de friction sont liés à l'écoulement aérien aux endroits où le conduit vocal présente un rétrécissement suffisant. Ils ont leur spectre propre (bruit de jet), coloré par la partie du conduit vocal qui se trouve en aval du point de constriction. On peut donc prévoir qu'un bruit de friction produit au niveau des lèvres aura une enveloppe spectrale plus simple qu'un bruit produit dans une région postérieure : c'est ce que l'on observe pour le /f/, bruit à bande large et très faible niveau, le /s/ qui présente une coloration bien marquée par une bande principale, et le /ʃ/ qui requiert deux bandes de bruit (fig. IV-6).

Les sons fricatifs peuvent être "voisés" ou "non voisés", c'est-à-dire accompagnés ou non de la vibration des cordes vocales. Dans le cas des sons voisés, en régime permanent, on pourrait penser que l'intensité du spectre de raies est liée à celle du bruit, l'air qui produit les deux phénomènes ayant la même origine. Ce serait oublier que l'ouverture de la glotte permet de réduire l'amplitude et la richesse harmonique de la vibration, tout en augmentant le débit d'air, donc l'amplitude du bruit.

Les bruits d'explosion proviennent d'une occlusion momentanée du conduit vocal,en un point donné, variable, suivie d'une brusque ouverture. Il se produit donc une compression d'air en amont de la zone obturée, et lors de l'ouverture un échelon de pression est appliqué à la partie aval du conduit. Ceci du moins pour les plosives non voisées ou "sourdes" /p/, /t/, /k/; pour les plosives voisées (ou "sonores" /b/, /d/, /g/) les cordes vocales continuent à fonctionner pendant la compression avec, néanmoins, un spectre moins étendu : seuls subsistent les premiers harmoniques. Le fonctionnement implique un débit aérien, obtenu au prix d'une légère fuite d'air par les narines, fuite dosée par le degré de fermeture du voile du palais. Pour que la compression (donc l'explosion) reste comparable à celle des plosives non voisées, il faut augmenter la pression sous-glottique. Ceci est valable en régime permanent ou lors d'une articulation lente; mais dans la parole courante, l'occlusion étant généralement de courte durée, on ferme complètement le voile du palais, et les cordes vocales continuent à fonctionner tant que la pression dans le pharynx est inférieure à la pression sous-glottique. Pratiquement le temps nécessaire à l'égalisation des pressions (plusieurs centaines de millisecondes) est largement suffisant pour permettre le fonctionnement des cordes vocales pendant

...../

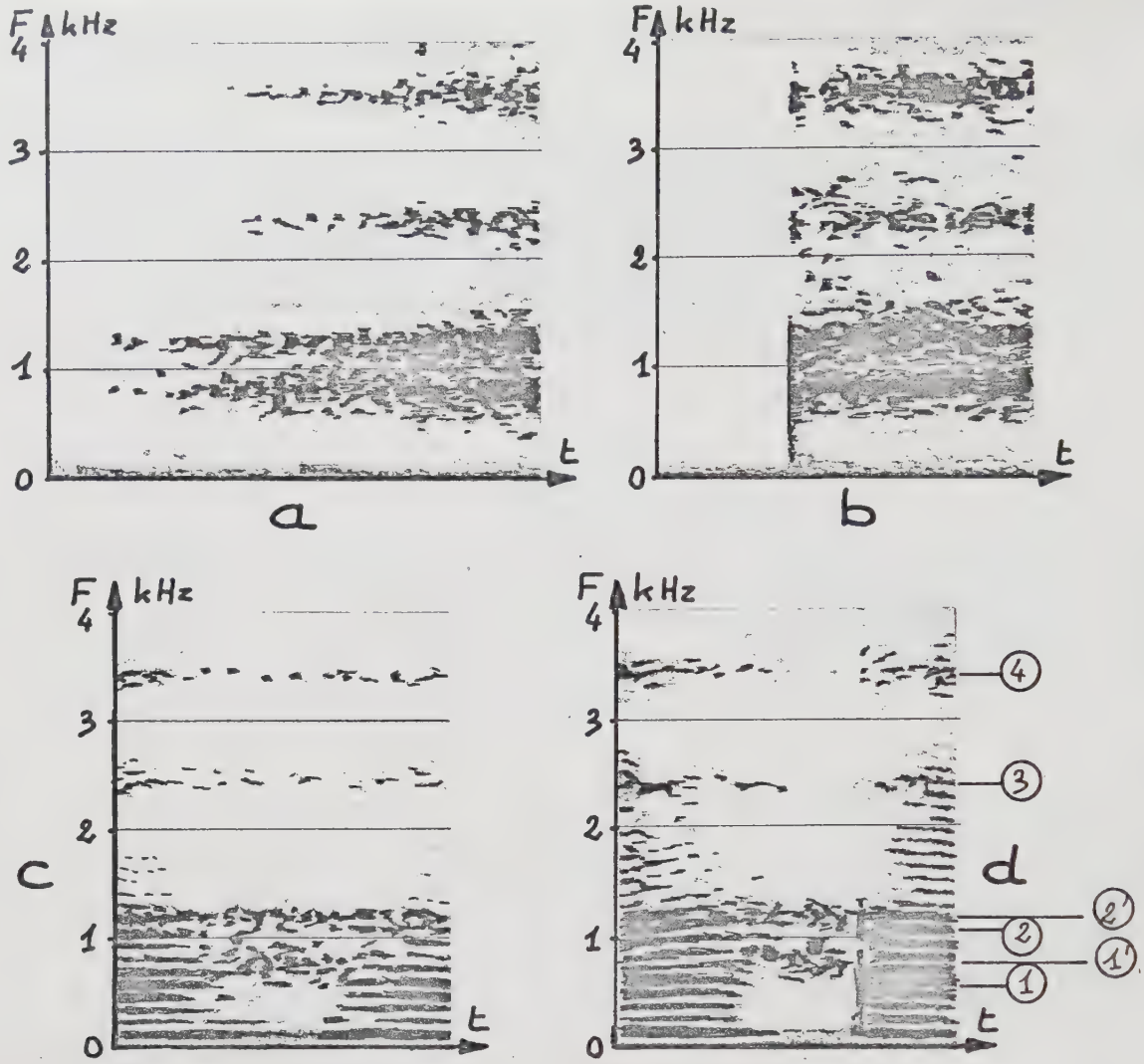


Fig IV -5 Mise en évidence de l'attaque glottique. a) Attaque progressive de la voyelle /a/, en voix chuchotée, locuteur masculin. b) Attaque glottique. c) Transition progressive voisé-chuchoté-voisé, sur la même voyelle. d) La même transition, comportant une attaque glottique. Sur cette dernière figure, on peut observer le décalage des deux premiers formants en voix voisée (1 et 2) et en voix chuchotée (1' et 2').

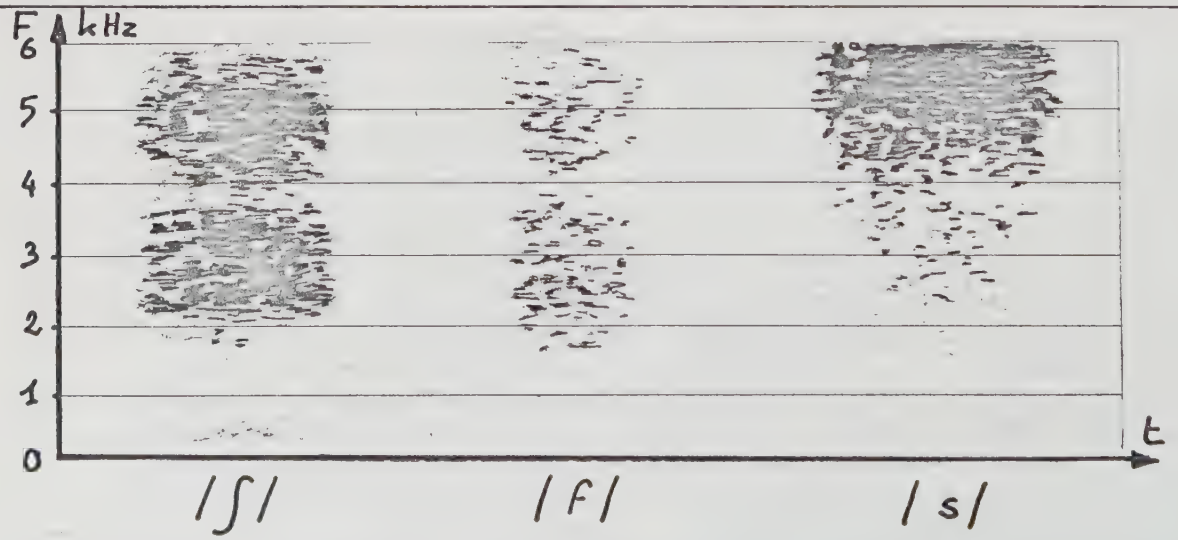


Fig IV -6 Consonnes fricatives sourdes, locuteur masculin.

sont normalement non jointives ou peu tendues; en tout état de cause, leur mise en vibration devrait être différée et progressive, ce qui n'est pas toujours observé. C'est ici qu'interviennent, selon E. LEIPP (bib. IV-1), les bandes ventriculaires, qui jouent un rôle comparable au "grain d'orge" de l'embouchure de trompette. Plus exactement, pour obtenir une mise en vibration rapide des cordes vocales, il faut deux opérations distinctes (fig. IV-4) : dans un premier temps, les bandes ventriculaires sont jointives, la longueur, l'épaisseur et l'ouverture des cordes vocales sont ajustées pour produire ultérieurement le spectre désiré, la pression sous-glottique, qui règne également dans les ventricules de MORGAGNI, est amenée à une valeur en excès sur la pression buccale (qui elle-même peut être supérieure à la pression extérieure si les lèvres et le voile du palais sont fermés). Dans un deuxième temps les bandes ventriculaires d'écartent brusquement, et la différence entre la pression sous-glottique et la pression buccale permet un écoulement localisé au niveau du larynx et une mise en vibration instantanée des cordes vocales.

Si un tel dispositif de surpression locale instantanée n'existait pas, la mise en vibration des cordes vocales serait toujours soit progressive, soit soumise à la fermeture de la bouche et du voile du palais, soit encore instantanée, mais à partir d'un état de fermeture totale des cordes vocales ce qui entraînerait un régime "voix criée" (plus intense, plus aiguë et plus riche) lors de chaque passage non voisé-voisé.

La commande de ce dispositif est volontaire : on peut obtenir différentes attaques d'un son vocalique, même en voix chuchotée, ainsi que différents types de transitions chuchoté-voisé (fig. IV-5). Notons que l'émission d'une voyelle en début de phrase s'accompagne généralement d'une "attaque glottique" voisine du son /k/, mais ignorée par la langue française.

IV.1.2.2. Fonctionnement du conduit vocal

Le conduit vocal a un double rôle de transmission d'une part, de génération de bruits et d'explosions, d'autre part.

a) Transmission acoustique dans le conduit vocal

Une cavité fonctionne en résonateur lorsque la pression acoustique est uniforme en chaque point. Cette condition est vérifiée lorsque les dimensions sont faibles devant la longueur d'onde, et lorsque les orifices ont une faible section. Il est clair, d'après la fig. IV-3, que les cavités de l'appareil phonatoire sont mal délimitées, ainsi que les orifices de liaison. D'autre part, les dimensions du conduit se comptent en centimètres, (longueur moyenne d'un conduit vocal masculin : 17 cm) et les longueurs d'onde correspondant aux formants de la parole (500-5000 Hz) varient entre 70 et 7 centimètres. Il est donc tout à fait faux d'expliquer le fonctionnement du conduit vocal au moyen de résonateurs localisés.

Il est plus raisonnable de le considérer comme un tube acoustique de section variable en fonction du temps et de la distance au larynx, et de longueur totale légèrement variable en fonction de la montée du larynx et de l'avancée des lèvres, les fosses nasales constituant une double dérivation au niveau du pharynx. Il faut également tenir compte des pertes de charge et de l'amortissement considérable des parois internes de la cavité, qui sont des tissus acoustiquement très absorbants. On considère alors la transmission d'une onde plane

...../

cordes vocales varie, en moyenne, du simple au double.

IV.1.2. Fonctionnement de l'appareil phonatoire

IV.1.2.1. Fonctionnement du larynx

Le larynx est un générateur sonore dont le spectre et le niveau peuvent être largement ajustés par le locuteur. Le son est produit par la vibration des cordes vocales, vibration complexe en trois dimensions, qui s'apparente à une oscillation de relaxation et comporte donc un grand nombre d'harmoniques (nombre variable selon les individus et les conditions de la phonation). Nous admettrons en première approximation les points suivants :

- La tension, la longueur et l'épaisseur des cordes vocales déterminent le fondamental du son émis.

- L'ouverture de la glotte détermine l'amortissement et la proportion de "bruit" du spectre : complètement fermées, les cordes vocales vibrent à la manière d'une anche double (voix "voisée"); complètement ouvertes, le son produit est un bruit d'écoulement (voix chuchotée); en fait, il existe une quantité de positions intermédiaires pour lesquelles le spectre de raies s'amortit plus ou moins, et s'ajoute au bruit d'écoulement.

- La pression sous-glottique est, toutes choses égales par ailleurs, responsable du niveau sonore.

Il est bien évident que tous ces paramètres sont en interaction constante. La voix "criée", par exemple, correspond au besoin d'émettre un fort niveau sonore. Pour cela, le locuteur augmente la pression sous-glottique, ferme la glotte, tend et enraidit les cordes au maximum. Il en résulte effectivement une forte pression acoustique, mais aussi une augmentation de la fréquence fondamentale et un enrichissement du spectre en harmonique, ce qui contribue à l'augmentation du niveau subjectif (cf zone sensible). Le champ de liberté acoustique du larynx connaît donc certaines limites d'origine purement mécanique, dont nous pouvons mentionner quelques aspects :

- La voix chuchotée est toujours de faible niveau; sa dynamique est faible par rapport à celle de la voix voisée. En voix chuchotée la glotte est ouverte et les cordes vocales ne vibrent pas; pour augmenter le niveau de bruit, il faut augmenter la pression sous-glottique; mais alors les cordes vocales entrent en vibration, et la seule ressource est de les ouvrir au maximum et de les enraidir; ce faisant, le débit d'air devient prohibitif et le locuteur s'épuise rapidement, sans que la pression sous-glottique augmente beaucoup.

- La voix criée correspond toujours à une tessiture haute du locuteur : on ne peut pas crier dans le registre grave, pour les raisons vues plus haut.

L'aspect temporel de mise en route du régime vibratoire est extrêmement intéressant. En voix criée le démarrage est instantané, cer les cordes vocales sont préalablement tendues et fermées; et c'est un excès de pression sous-glottique qui détermine le premier cycle de relaxation. En voix normale par contre, il se produit - ou devrait se produire - un léger écoulement aérien avant la mise en vibration des cordes vocales, car celles-ci

...../

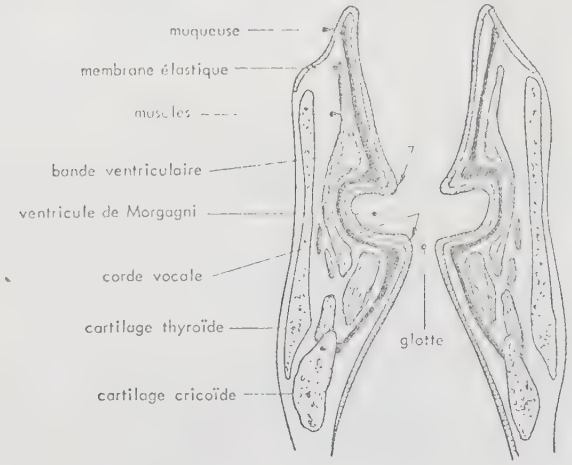


Fig IV-1

Coupe verticale du larynx (d'après E.LEIPP).

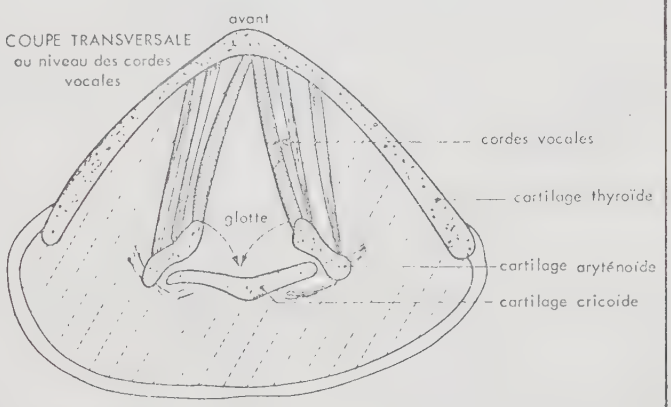


Fig IV-2

Coupe horizontale du larynx, au niveau des cordes vocales (d'après E.LEIPP).

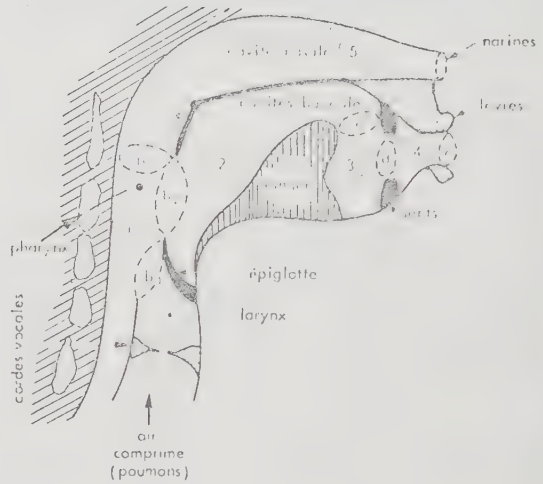


Fig IV-3

Coupe sagittale du conduit vocal (d'après E.LEIPP).

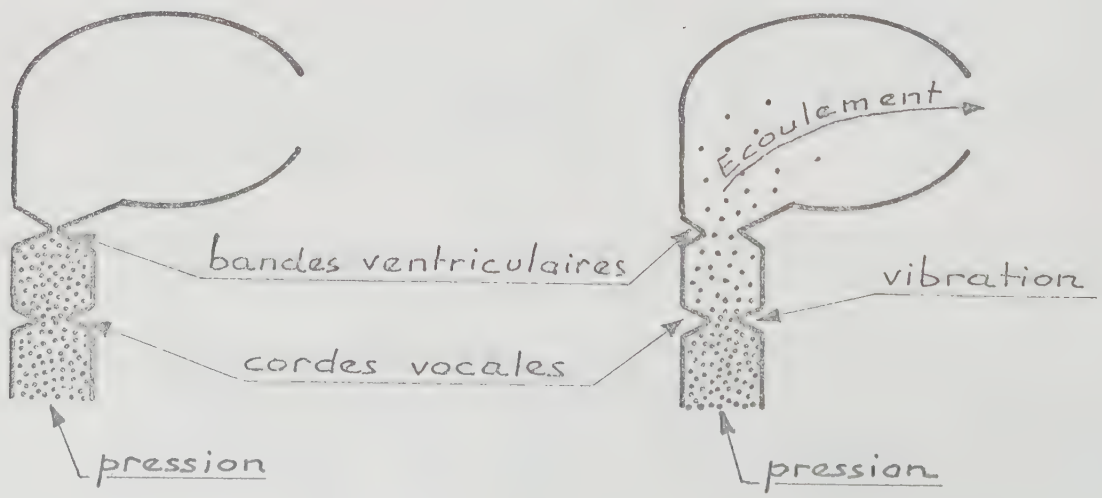


Fig IV-4 Rôle des bandes ventriculaires dans la mise en vibration des cordes vocales.

IV.1.1.1. Le larynx (fig. IV-1)

Le larynx est à lui seul un organe d'une grande complexité mécanique. Il comprend divers cartilages mobiles (aryténoïdes et thyroïde principalement), reliés entre eux par des ligaments et des muscles striés. La partie interne est garnie d'une muqueuse formant deux paires de bourrelets membrano-musculaires (cordes vocales et bandes ventriculaires) séparés par une double cavité (ventricules de MORGAGNI). Les cordes vocales et les bandes ventriculaires ont une extrémité antérieure commune fixée sur le cartilage thyroïde, qui peut basculer vers l'avant et tendre plus ou moins l'ensemble (fig. IV-2). Les extrémités postérieures des cordes vocales et des bandes ventriculaires sont fixées sur les deux cartilages aryténoïdes, qui possèdent plusieurs degrés de liberté. Le résultat de cette articulation complexe est le suivant : l'ouverture des cordes vocales (glotte), qui présente la forme d'un triangle isocèle, peut être réglée en largeur et en hauteur, ainsi que l'ouverture des bandes ventriculaires.

La tension et l'écartement des cordes vocales d'une part, des bandes ventriculaires de l'autre, sont donc réglables indépendamment. Il est possible également d'épaissir les cordes vocales en contractant les muscles qui les sous-tendent et même de déplacer l'ensemble du larynx vers le haut dans une large mesure (plusieurs centimètres), ce mouvement servant avant tout à la déglutition (l'épiglotte vient alors obturer la glotte).

IV.1.1.2. Le conduit vocal (fig. IV-3)

La section sagittale du conduit vocal montre un profil extrêmement variable selon la position des organes qui le délimitent. Il est bien difficile de distinguer précisément les cavités les unes des autres, d'autant que des coupes transversales montreraient une section très variable depuis la glotte jusqu'aux lèvres. Cependant, on distingue usuellement quatre cavités principales : le pharynx, partie postérieure du conduit vocal, de volume très variable selon l'élévation du larynx et la position de la langue; les fosses nasales, mises en communication avec le pharynx par l'abaissement du voile du palais; les cavités buccales, formées par la langue et la voûte palatale, cavités qui se réduisent généralement à un seul "tube" de forme et de section complexes; enfin l'ajutage lippal qui constitue la partie antérieure du conduit.

Chacun des organes en jeu possède un champ de liberté considérable, mis à profit par l'organisme au moyen de très nombreux muscles. La langue, par exemple, dispose à elle seule de 17 muscles (8 pairs et 1 impair); la mâchoire inférieure comporte 10 muscles; les lèvres en comportent également une dizaine, etc... On peut affirmer qu'une cinquantaine de muscles au moins participent synergiquement à la phonation.

A la complexité musculaire répond une égale complexité nerveuse : sur les neuf paires de nerfs crâniens comportant des fibres motrices, six sont concernées par le système phonatoire.

Il convient de remarquer également la très grande variété possible dans les dimensions des organes phonatoires, d'un individu à l'autre. Entre l'adulte et l'enfant qui commence à parler (2 ans) le rapport des dimensions linéaires, bien que variable selon l'organe considéré, est de l'ordre de 2 ; c'est-à-dire que le rapport des sections est de l'ordre de 4, et le rapport des volumes de l'ordre de 8. De même, entre deux individus de même corpulence les différences locales peuvent être considérables. Entre une femme et un homme la longueur des

...../

IV - STRUCTURE PHONETIQUE ET ACOUSTIQUE DE LA PAROLE

Le message parlé s'est élaboré en fonction des caractéristiques de l'émetteur et du récepteur. Nous avons examiné dans le chapitre précédent quelques aspects fonctionnels du récepteur, c'est-à-dire de l'oreille. Nous pensons que les contraintes imposées par l'audition sont primordiales dans la genèse du message parlé, et que celles provenant de la phonation n'interviennent qu'en second lieu; en d'autres termes, qu'il est plus facile de faire une machine à parler qu'une machine à écouter.

Il est néanmoins indispensable de décrire dans ses grandes lignes l'appareil phonatoire humain, ne serait-ce que pour souligner sa complexité et sa variabilité d'un individu à l'autre. Le mécanisme supposé de la phonation a de tous temps servi de base au système descriptif de la parole : jusqu'à ces dernières années, la phonétique ne s'intéressait guère qu'aux positions statiques du système phonatoire, qui correspondent à l'émission de sons permanents. La nécessité d'étudier les transitions phonétiques n'est apparue que vers 1950, lorsque DELATTRE et ses collaborateurs ont entrepris la synthèse de la parole. Ces travaux sont malheureusement empreints d'une systématisation excessive : la théorie du "locus", qui tente de définir chaque consonne par le point de convergence des formants vocaliques, n'est pas vérifiée dans les faits.

Pour aller plus loin, il faut considérer la parole comme une structure acoustique définie dans le plan temps fréquence. Cette structure, ou Forme, possède des propriétés d'anamorphose, d'ordre proche ou lointain, d'émergence sur un fond, etc... Comme il s'agit d'une forme faible ou articulée, on peut, de manière assez artificielle, la considérer comme une simple juxtaposition d'éléments, qui doit cependant respecter la continuité de fonctionnement des organes phonatoires : ces considérations mènent à la définition du phonatome, ou élément de phonation, dont nous montrerons qu'il constitue une unité significative de la parole.

La notion de phonatome possède diverses conséquences pratiques, dans les domaines de la synthèse, de la reconnaissance, et de la linguistique quantitative, dont certaines (contraintes phonétiques, mots aléatoires) seront examinées dans ce chapitre.

IV.1. LA PHONATION (bib. IV-1, III-2, II-6)

IV.1.1. Description sommaire du système phonatoire

L'appareil phonatoire se compose de deux parties anatomiquement distinctes. Les poumons et le larynx, partie supérieure de la trachée artère, constituent l'essentiel du générateur sonore; l'ensemble des cavités et ajutages mobiles délimités par le rhino-pharynx, le voile et la voûte du palais, la langue, les maxillaires inférieur et supérieur, les joues, les dents, les lèvres est appelé le conduit (ou "tractus") vocal.

...../

Un modèle de l'audition particulièrement intéressant a été donné par A. MOLES (bib. III-21).

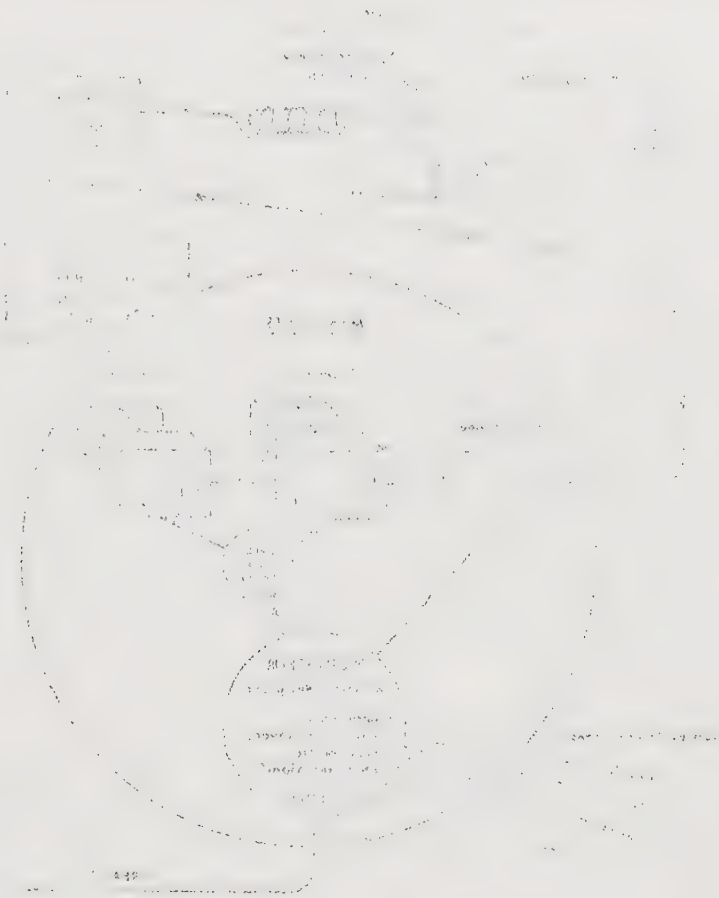
Les études que nous décrirons par la suite sont sous-tendues par le modèle de l'audition proposé par E. LEIPP (bib. III-8,I-2) : La fonction auditive est scindée en deux parties (fig. III-14). La première comporte un compresseur de dynamique par rétroaction instantanée et une analyse temps-fréquence du message acoustique. Cette analyse est symbolique : dans la réalité, la dimension de fréquence peut être codée tout-à-fait différemment et sans souci de précision. Associée à une mémoire en boucle de quelques secondes (mémoire phosphorescente), l'analyse fournit une certaine représentation globale de l'objet sonore, dans laquelle les phases sont éliminées et la dynamique réduite.

Cet objet, qui relève de la Théorie des Formes, est codé au moyen d'impulsions et transmis au centre de calcul. Celui-ci est composé d'un organe de calcul susceptible de comparer et de classer les objets, et de deux mémoires, l'une à court terme, l'autre à long terme. Le centre de calcul traite également d'autres messages sensoriels, assure la gestion de l'ensemble et prend des décisions.

Un tel modèle permet de donner une structure à certaines activités psychiques entendre, écouter, dormir, rêver etc... correspondent à certains types de fonctionnement de cet ensemble. En particulier, les différences de compréhension d'un même message (ou ensemble d'objets sonores codés) selon sa prévisibilité sont illustrées ici par le fait que le message-référence a été ou non transmis au préalable de la mémoire permanente à l'organe de calcul. Dans un cas le message à identifier donne lieu à une recherche exhaustive en mémoire; dans l'autre le domaine de recherche a été fortement délimité par une perception précédente (auditive, visuelle, ou autre). Et s'il est bien vrai que la compréhension est un pari sur le message à venir, il est vrai également que ce pari s'effectue avec plus de chances de succès si les informations sensorielles passées ont permis de restreindre le champ des recherches en mémoire.

III.5. CONCLUSION

Nous avons examiné dans ce chapitre divers aspects de l'audition et de la perception auditive - La physiologie ne nous a donné que des observations fragmentaires : l'aspect "téléphonique" des connexions neuroniques n'est pas d'un grand secours si on ne sait pas quel usage est fait de l'information à tous les niveaux. La psychophysique de l'audition est obligée, pour rester objective, de se placer dans des conditions expérimentales fort éloignées de la réalité. Les conceptions globales qu'apportent la théorie de la Forme et le structuralisme sont plus conformes à l'observation courante. Mais les théories ne fournissent qu'un cadre de pensée, alors que les modèles et les techniques de simulation fournissent des outils. Pour résoudre le problème qui nous intéresse - celui des structures acoustiques de la parole et de leur reconnaissance - nous choisirons les techniques expérimentales (synthèse) ou pseudo-expérimentales (modèles). Avant d'aborder ces aspects plus concrets du problème, il nous faut encore examiner le mécanisme de la phonation et la description phonétique de la parole.



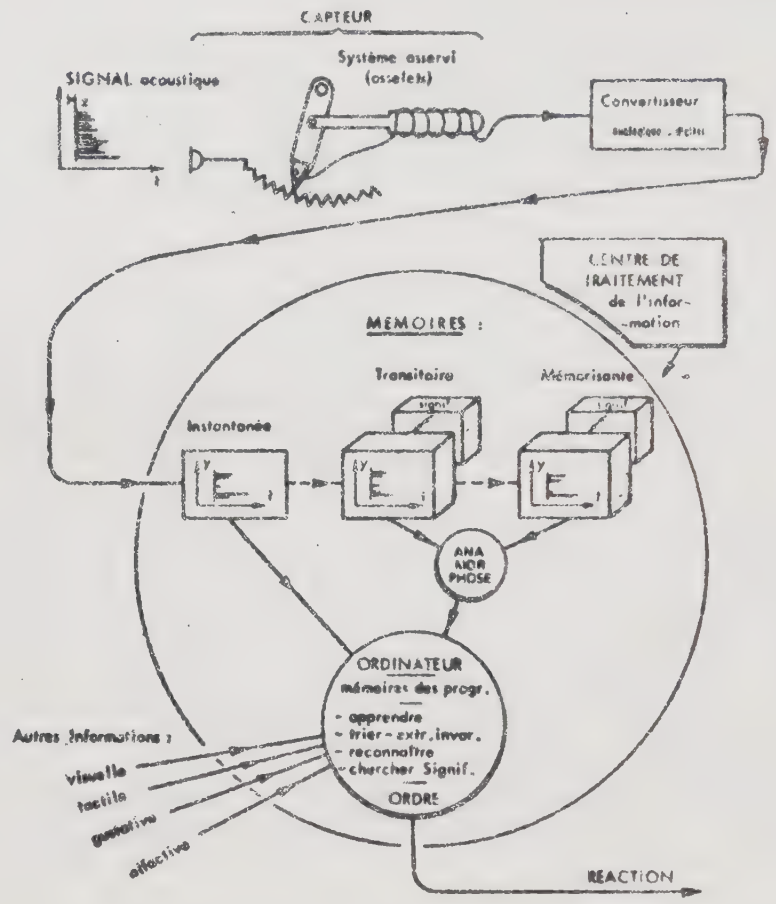


Fig III-14 Modèle fonctionnel de l'audition (d'après E.LEIPP).

La forme acoustique est captée par l'oreille, qui comporte un asservissement en niveau et en timbre, puis codée dans la cochlée sous forme d'un ensemble d' impulsions. Cette configuration possède une certaine rémanence (mémoire instantanée) qui en permet un premier examen par le centre de traitement. Si c'est utile, la configuration incidente est rangée dans un fichier provisoire (mémoire transitoire), afin de servir elle-même de référence ultérieure, moyennant anamorphose. Ce fichier est en évolution permanente: les fiches les plus utiles et les plus fréquentes sont classées dans un fichier définitif (mémoire mémorisante).

zones dans lesquelles on peut trouver chacun des formants. Nous voyons ainsi que le second formant est le plus "mobile", et surtout le plus caractéristique : pour une même valeur du 2° formant, on n'a jamais le choix entre plus de 3 voyelles, ce nombre allant jusqu'à 8 pour le premier formant et 9 pour le troisième. De plus, l'étendue du 2° formant coïncide à peu près avec la zone sensible de l'oreille, pour un ensemble de locuteurs hommes, femmes et enfants.

IV.2.2.5. Les traits distinctifs des voyelles

Nous avons évoqué plus haut certaines parentés entre voyelles. Ces ressemblances et oppositions sont bien connues des linguistes, qui les ont étudiées systématiquement, ainsi que des orthophonistes, qui les utilisent en rééducation de la parole.

JAKOBSON a mis en évidence la hiérarchie entre les distinctions phonologiques, en cherchant à définir chaque phonème par un ensemble limité de qualificatifs nommés traits distinctifs. Ainsi les voyelles françaises prennent-elles place, d'après B. MALMBERG (bib. III-16), dans un schéma arborescent comportant 5 niveaux (fig. IV-13) qui correspondent aux qualificatifs de "nasalité", "compacité", "gravité", "complexité", et "labialisation" ("flatting"). Les traits distinctifs peuvent être ou non "pertinents" pour un phonème donné : pour /a/ et /ɑ/, la complexité et la labialisation ne sont pas des traits pertinents. Le schéma indique, par exemple, que les voyelles /i/ et /y/ ne diffèrent que par un trait distinctif (labialisation), alors que /a/ et /ɔ̃/ diffèrent par deux traits (nasalité et gravité).

On conçoit l'intérêt d'un tel classement pour la description phonologique. L'étude du langage enfantin montre d'ailleurs que les distinctions entre phonèmes voisins sont les plus tardivement acquises (bib. IV-5). Il est donc possible de déduire de cette classification un système minimal, dans lequel les distinctions trop subtiles ont disparu (fig. IV-14).

Cependant, les traits distinctifs, établis en fonction d'observations articulatoires, sont trop mal définis sur le plan acoustique pour être d'une grande utilité en reconnaissance automatique. Il est bien difficile, par exemple, de trouver un caractère acoustique simple représentant la nasalité ou la labialisation ! .... Nous retiendrons donc que la théorie des traits distinctifs fournit une classification intéressante des voyelles isolées, mais part d'hypothèses trop atomistes pour s'appliquer à une étude d'ensemble des formes de la parole.

IV.2.3. Les consonnes

A part les fricatives /s/, /z/, /ʃ/, /ʒ/, /f/, /v/, les consonnes n'ont pas d'existence autonome. Elles sont caractérisées par un mouvement phonatoire, c'est-à-dire par une transition phonétique, qui prend généralement une voyelle pour point de départ ou d'arrivée. Il est bien évident que tout mouvement du conduit vocal entraîne une modification de ses résonances et, partant, que toute consonne s'accompagne d'une évolution des formants vocaliques.

DELATTRE et LIBERMANN (bib. IV-6) ont tenté de systématiser cette constatation par la théorie du Locus. Selon celle-ci, toute consonne est définie par la fréquence du point de convergence des formants vocaliques. Cette systématisation était

...../

évidemment excessive : pour appliquer la théorie il a fallu définir plusieurs loci pour certaines consonnes (/k/ par exemple). La théorie était d'ailleurs sans portée pratique pour l'analyse, car l'imprécision inhérente à la mesure de fréquence des formants est telle, en parole normale, qu'il n'est guère possible de définir leur point de rencontre.

Néanmoins, ces expérimentations, effectuées sur le synthétiseur "play-back" des laboratoires HASKINS en 1950 (fig. V-7) ont mis l'accent sur l'importance des transitions phonétiques. Depuis cette époque, phonéticiens et acousticiens ont cherché à définir explicitement les lois de variation des formants dans la transition d'un phonème à l'autre. Malgré cette évolution, la conception atomiste de la parole demeure, et la notion de phonème est prépondérante. Pour aller plus loin dans notre étude, il est nécessaire de montrer que les structures acoustiques de la parole sont régies par leur perception et, à travers cette dernière, par la théorie de la Forme.

IV.3. STRUCTURE ACOUSTIQUE DE LA PAROLE

Nous nous efforcerons de faire abstraction, dans ce paragraphe, de l'aspect articulatoire de la parole, pour diverses raisons déjà mentionnées : les paramètres articulatoires sont trop mal connus, leur variabilité d'un individu à l'autre est considérable, et surtout un même message acoustique produit par deux locuteurs différents n'implique nullement un fonctionnement identique de leurs appareils phonatoires. Nous prendrons pour seule preuve de ce dernier point l'extraordinaire aptitude du mainate (merle des Indes) à imiter la voix humaine (bib. IV-7) : après conditionnement, cet oiseau est capable de répéter des phrases complètes - intonation, rythme, accentuation et squelette sémantique compris - avec un tel degré de perfection que la plupart des auditeurs non prévenus sont incapables de discerner l'origine animale du message. Seul, semble-t-il, le timbre change : la "voix" du mainate possède moins d'énergie dans le registre grave, et il en résulte un timbre plus "nasillard" que celui de la voix humaine, ce terme étant évidemment impropre puisque le mainate n'a ni nez, ni fosses nasales. Cette observation fondamentale renforce notre scepticisme quant à la pertinence des paramètres articulatoires dans le domaine acoustique, l'appareil phonatoire du mainate étant évidemment fort différent de celui de l'homme tant dans son anatomie (pas de langue, pas de dents ni de lèvres) que dans son fonctionnement.

Pour mieux accentuer le point de vue "non-articulatoire" que nous venons de définir, nous présenterons en premier lieu les caractéristiques "externes" du message parlé, celles que l'on doit maîtriser pour conserver son intelligibilité dans la transmission électroacoustique par exemple. Nous montrerons ensuite que l'on peut mettre en évidence le squelette sémantique, qui transporte la signification du message et relève de la Théorie de la Forme. Le squelette sémantique se présente généralement comme une forme articulée, ou forme faible, que l'on peut segmenter en formes élémentaires ou phonatomes, en examinant son autocorrélation à court terme. La courbe de stabilité du squelette sémantique nous permettra de préciser la notion de "netteté" de la parole, notion fondamentale dans tous les problèmes touchant à l'intelligibilité. Nous examinerons enfin l'application au message parlé de la notion d'anamorphose : anamorphoses fréquentielles et temporelles du squelette sémantique, anamorphoses plus générales de la voix, qui déterminent les champs de liberté indispensables à l'adaptation à des locuteurs variés et à la communication sur le plan esthétique.

...../

IV.3.1. Caractéristiques externes du message parlé

Il faut distinguer deux types de transmission de la parole : celle ("haute fidélité") qui cherche à restituer toute l'information physique du message original, et celle (type téléphone) qui vise en premier lieu à la transmission de l'information sémantique, avec un minimum d'éléments esthétiques ou diagnostiques (intonation, rythme). A ces deux conceptions correspondent des appareillages différents ou des architectures différentes (salles de concert, amphithéâtres). En général, on considère le second aspect comme le plus important; nous verrons que le premier n'est pas à dédaigner, en tant qu'outil de test.

IV.3.1.1. Bande passante

On peut admettre qu'une transmission de parole est très fidèle si les fréquences comprises entre 100 et 10 000 Hz sont correctement rendues; mais le téléphone se contente de 300-3400 Hz. Cette bande peut s'avérer un peu étroite si l'on veut percevoir certaines nuances du registre aigu (sons isolés /s/, /f/ etc... qui nécessitent une épellation codée : "Suzanne", "François", etc...), tout particulièrement pour les voix de femmes et d'enfants. Pratiquement une bande de l'ordre de 500-5000 Hz est tout-à-fait suffisante.

Nous avons effectué certains tests, classiques en téléphonométrie, pour déterminer la contribution de différentes bandes de fréquence à l'intelligibilité des mots courants. Ces mots (listes de FOURNIER) ont été lus et enregistrés par deux locuteurs, masculin et féminin, puis filtrés par bande d'octave et écoutés par une trentaine d'auditeurs. Les résultats (taux de mots reconnus globalement, fig. IV-15) montrent que les bandes 1000-2000 Hz, puis 2000-4000 Hz, sont les plus importantes. Nous rejoignons par là la notion de zone sensible évoquée au chapitre précédent. Donc, pour tirer parti d'une bande passante trop étroite, il faut centrer celle-ci sur 2000 Hz environ. On peut vérifier cette observation par l'étude acoustique de certains systèmes de reproduction sonore réglés empiriquement (mini-transistors, poupées parlantes, etc...)

IV.3.1.2. Dynamique

Le rapport signal-bruit de la parole est une quantité éminemment variable selon plusieurs facteurs :

a) La "voix normale" du locuteur peut être plus ou moins chargée de bruit, bruit continu dû à l'écoulement aérien au niveau de la glotte, qui diminue l'émergence des formes de la parole, donc altère la netteté; ou clics parasites dus à la viscosité de la salive.

b) Le type de voix adoptée est essentiel : la dynamique de la voix chuchotée est de l'ordre de 15 dB, alors que la voix criée peut atteindre 80 dB dans les mêmes conditions. Ordinairement, on peut compter sur une quarantaine de dB.

c) Le niveau du bruit ambiant ne doit pas être distingué de sa structure acoustique. L'électroacoustique ne prend en compte que le bruit blanc gaussien à large bande, qui se trouve être très rare dans la nature, même industrialisée : on ne le rencontre guère que dans les appareillages électroniques et

...../

à proximité des chutes d'eau. En réalité, chaque bruit a sa structure propre, et son effet sur le message parlé relève de la théorie de la Forme plus que du niveau en décibels, bruts ou pondérés. La notion de rapport signal-bruit doit céder la place à celle d'émergence d'une forme sonore sur un fond (voir E. LEIPP, bib. III-14 et IV-8).

L'expérimentation relative à la dynamique de la parole fournit un résultat intéressant, sinon surprenant : la dynamique nécessaire pour transmettre l'information sémantique est très réduite. Dans le cas où le bruit masquant est un bruit blanc, on constate même que le message n'est pas totalement détruit lorsque le niveau de bruit excède le niveau du message d'une dizaine de dB (en voix voisée normale). Ceci permet de comprendre pourquoi le téléphone fonctionne malgré sa faible dynamique (10 à 20 dB).

En électroacoustique, on cherche à réduire la dynamique des messages car celle des moyens d'enregistrement est limitée. Cette opération ne présente aucun inconvénient pour la parole, et peut même être effectuée automatiquement au moyen de "compresseurs" électroniques. Un système particulièrement efficace est celui de l'enregistreur NAGRA IV, qui évite la remontée de bruit pendant les "silences" au moyen d'une double constante de temps. On notera l'analogie avec la commande musculaire de l'oreille moyenne (E. LEIPP, bib. III-3) et la notion de retard variable exposée plus haut (au chapitre III), qui nous mènera à simuler l'adaptation en niveau lors du processus de reconnaissance automatique.

IV.3.1.3. Distorsions

Il serait très intéressant d'étudier les distorsions du message parlé comme cela été effectué pour les sons musicaux (MOLES, bib. IV-9). Mais les mauvais traitements que l'on peut infliger au signal parlé étant en nombre infini, nous nous limiterons à quelques cas-types.

a) Altérations de la courbe de réponse

Dans la mesure où la bande 500-5000 Hz est à peu près préservée, même avec des lacunes, le message reste intelligible. Les haut-parleurs ont très souvent une courbe de réponse extrêmement torturée, les inégalités étant masquées par un fort amortissement des appareils de mesure (bib. IV-10); cela n'empêche nullement une bonne transmission de l'information sémantique.

Les différentiations ou intégrations effectuées sur le signal f(t) entrent dans ce même type de distorsion, car elles se ramènent à des pondérations aiguës ou graves de 6 dB/octave. Des essais de R. COUNORD (bib. IV-11) il ressort que cinq différentiations successives n'altèrent pas l'intelligibilité mais que le timbre devient "mordant", "crispant". Par contre, au bout de trois intégrations l'intelligibilité tombe aux alentours de 20 %. Ce dernier point s'explique facilement par l'altération des fréquences situées dans la zone sensible : l'énergie des premiers harmoniques (aux alentours de 500 Hz) déterminant le niveau global; le gain a été réglé pour maintenir celui-ci constant; dans ces conditions, la zone sensible est affaiblie d'environ 36 dB, et l'information importante est perdue.

b) Distorsion harmonique et écrêtage

La distorsion harmonique et les phénomènes corollaires (intermodulation) n'affectent pratiquement ni l'intelligibilité ni le timbre de la parole. Ces

..../

distorsions se traduisent par la création d'harmoniques pendant les périodes voisées, et de bruit pendant les périodes non voisées. Il s'ensuit un simple renforcement de ces bruits et harmoniques, qui reste inaudible dans une large mesure.

Le phénomène ne devient sensible que lorsqu'il correspond à un véritable écrêtage; le timbre devient alors plus nasillard et un bruit caractéristique ("buzz") se fait entendre sur les sifflantes. Mais l'intelligibilité est peu altérée.

c) Distorsions spectro-temporelles

A cette classe appartiennent les transpositions du message provenant d'une vitesse de reproduction différant de la vitesse d'enregistrement (disque ou magnétophone), ou les altérations dues à l'irrégularité de la vitesse de défilement (pleurage, scintillement).

Ici encore l'oreille est très tolérante : l'information diagnostique est altérée dès que la transposition atteint 5 à 10 %, mais le message peut être compris alors que la vitesse est multipliée ou divisée par 1,5, voire par 2 ou plus avec un auditeur entraîné. Il s'agit là d'une anamorphose combinée temps-fréquence, fort utilisée pour donner une voix insolite aux personnages des dessins animés.

Le pleurage, très gênant sur des notes musicales tenues, n'est pratiquement pas sensible sur la parole, car la hauteur du fondamental est extrêmement fluctuante.

d) Réverbération

Aussi faible soit-elle, la réverbération affecte la quasi-totalité des enregistrements. Dans un local donné, chaque composante sonore est accompagnée d'une traînée d'autant plus longue qu'il s'agit d'une composante plus intense. Ce phénomène contribue à donner une orientation au temps, bien visible sur les sonagrammes, et tend à détruire les formes de la parole. Nous verrons plus loin que la notion de netteté du message pourrait être fructueuse dans l'étude acoustique des salles de parole.

IV.3.1.4. Résistance de l'information sémantique

Les quelques remarques qui précèdent nous montrent deux aspects complémentaires du message parlé. L'aspect sémantique, d'une part, extrêmement résistant à toutes les altérations : altérations plus ou moins complexes portant sur les trois dimensions de niveau, temps ou fréquence ; l'aspect esthétique et diagnostique, d'autre part, qui représente en quelque sorte le champ de liberté du premier. La parole, sous l'aspect esthétique, constitue un test très sensible dans l'appréciation de la qualité des systèmes de transmission.

Cette dualité trouve son explication dans la disparité des débits d'information sémantique (quelques dizaines de bits par seconde) et esthétique (environ mille fois plus), qui entraîne une grande redondance sur le plan sémantique. Ceci confirme encore une fois l'idée selon laquelle le message parlé s'est élaboré en fonction des contraintes du milieu naturel, dont les effets s'apparentent aux altérations et distorsions que nous venons d'évoquer.

...../

IV.3.2. Squelette sémantique et phonatomes

IV.3.2.1. Mise en évidence du squelette sémantique

Examinons les sonagrammes de la fig. IV-16. Malgré leurs nombreuses différences de timbre, de rythme, d'intonation, de niveau, il subsiste une certaine ressemblance qui correspond à l'invariant sémantique ou squelette sémantique, schématisé sur la figure IV-17. Cette schématisation n'est pas nécessairement la meilleure possible, mais elle est convenable à partir du moment où l'on a vérifié son intelligibilité par synthèse (cf. chapitre V).

Il faut reconnaître que les caractères du squelette sémantique ne sont pas évidents sur tous les sonagrammes. Le caractère de voisement, par exemple, est complètement exclu du message chuchoté. Mais ce dernier présente deux aspects qui le rendent extrêmement précieux :

a) L'absence de spectre de raies entraîne une simplification notable de la structure acoustique.

b) La dynamique est réduite : la différence de niveau entre les éléments vocaliques et les éléments explosifs ou fricatifs est beaucoup plus faible qu'en voix normale ou criée.

Toutes les composantes du messages sont donc sur le même plan en voix chuchotée; l'importance des bruits et des clics est ainsi mise en évidence (et vérifiée par synthèse). Ce point est important dans la mesure où la phonétique acoustique classique l'a complètement négligé.

IV.3.2.2. Squelette sémantique et phonatomes

Le squelette sémantique est une superforme, au sens de la Gestalt-Theorie. Nous devons admettre, comme le veut l'expérience, qu'il est perçu d'abord comme un tout, surtout s'il est déjà partiellement ou totalement prévisible. Peu importe que son tracé ne soit pas très net, qu'il subisse des anamorphoses limitées, ou des variations d'un individu à l'autre.

Mais cette superforme est généralement articulée en formes élémentaires possédant une plus grande cohésion interne, que nous avons appelées "éléments phonétiques", et par la suite "phonatomes" (bib. III-13 et IV-14).

Les phonatomes, éléments de mouvement du conduit vocal, sont séparés par des états quasi-stationnaires qui par eux-mêmes ne convoient aucune information, si l'information est l'imprévisibilité c'est-à-dire le changement. On peut donc trouver les frontières des phonatomes en étudiant la stationnarité (ou stabilité) du squelette sémantique. Pour cela, on compare le squelette sémantique à un instant donné $(t + \frac{\tau}{2})$ à ce qu'il était à l'instant $(t - \frac{\tau}{2})$, le paramètre τ (distance d'ordre) étant choisi entre 20 et 50 ms selon la vitesse d'élocution. Le résultat de cette comparaison est porté selon une courbe $S(t, \tau)$ évoluant entre 0 (stabilité nulle) et 1 (stabilité totale), dont les maxima relatifs permettent un découpage en phonatomes. L'exemple donné ci-contre (fig. IV-17) n'est qu'un schéma de principe de la segmentation phonétique que nous utiliserons lors de la reconnaissance automatique.

...../

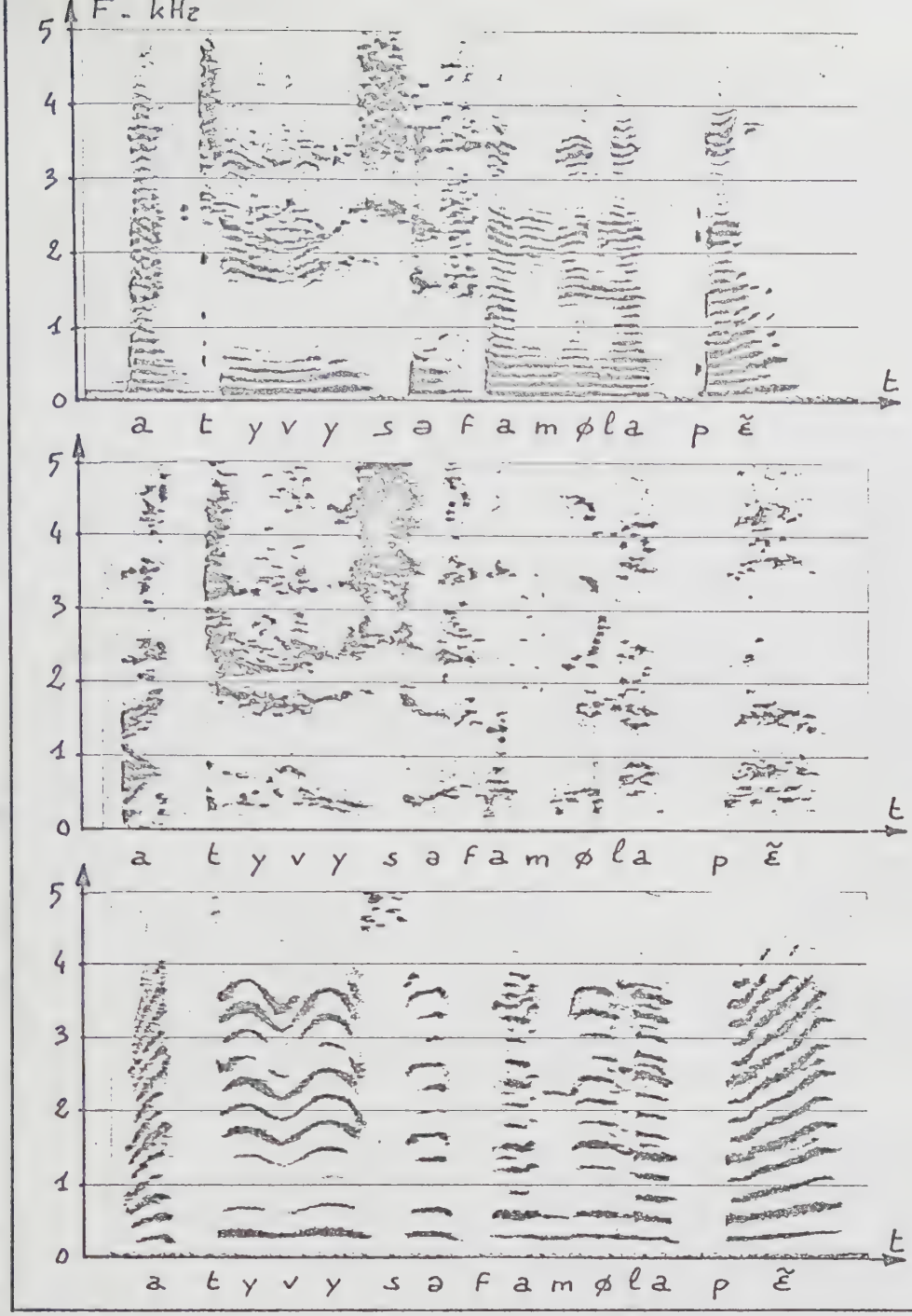


Fig IV-16

Divers sonagrammes, en bande étroite (ΔF =45 Hz) de la phrase "as-tu vu ce fameux lapin ?", prononcée par un locuteur masculin (JSL) en voix "normale", puis en voix chuchotée et en voix criée.

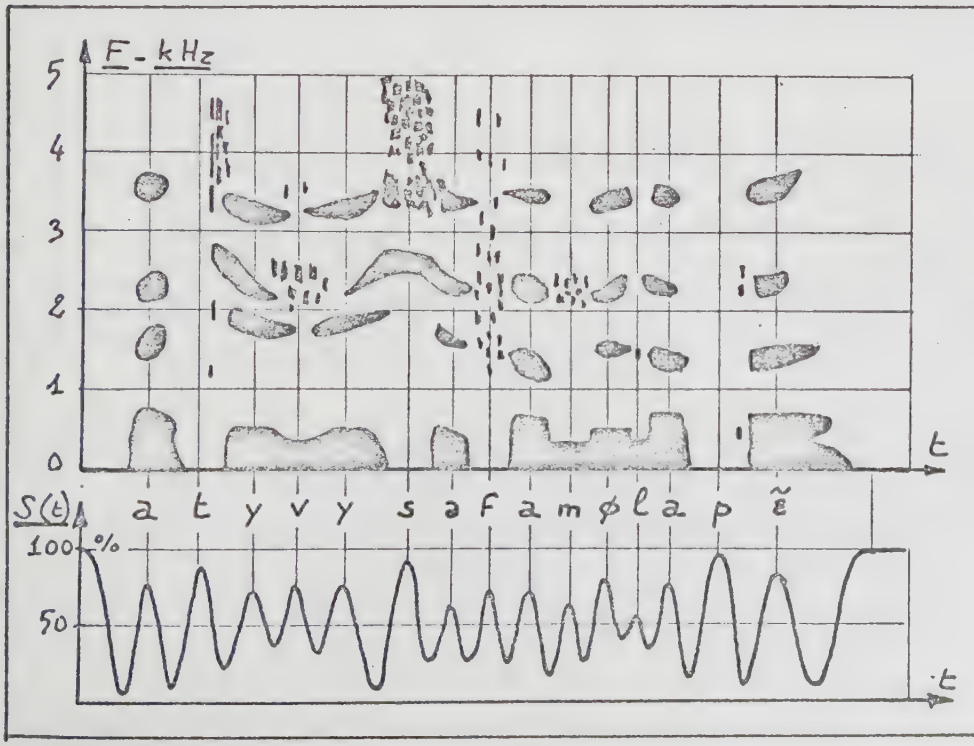


Fig IV-17

Squelette sémantique de la phrase précédente, accompagné de sa courbe de stabilité S(t), dont les extrema correspondent aux frontières des phonatomes.

Il faut bien répéter que le découpage en phonatomes, s'il est généralement observé dans la parole normale, ne correspond nullement à une nécessité fonctionnelle du message parlé. En parole rapide ou mal articulée, les états quasi-stationnaires sont souvent escamotés, et plusieurs phonatomes successifs peuvent se regrouper en une forme de longueur intermédiaire entre le phonatome et la superforme.

IV.3.2.3. Notion de netteté

Comme nous l'avons déjà constaté, il peut être plus ou moins difficile d'extraire le squelette sémantique du message réel (ou de sa représentation temps-fréquence). Cette difficulté tient à de nombreuses causes : mauvaise articulation, timbre moins riche du locuteur, bruit de fond, réverbération, voix de femmes ou d'enfants, etc... C'est un fait d'expérience que certaines voix sont plus "distinctes", "claires", "nettes", pour ne pas dire plus "intelligibles", que d'autres. Cette différence de netteté peut être mise en évidence par les sonagrammes, et se traduit également dans l'allure des courbes $S(t, \tau)$: plus le message est "net", et plus la courbe $S(t, \tau)$ est accidentée. En caractérisant ses variations soit par un histogramme et un écart-type, soit par une fonction de la valeur absolue de la dérivée, on peut envisager une étude quantitative de la netteté du message, qui trouverait des applications dans tous les problèmes d'intelligibilité.

IV.3.2.4. Phonèmes et phonatomes

La notion de phonème est abstraite. Il est utopique de rechercher des phonèmes dans le message acoustique, même aux instants de quasi-stationnarité mis en évidence par les maxima de la courbe $S(t, \tau)$, sauf peut-être dans le cas d'une parole lente et d'une prononciation particulière.

La notion de phonatome respecte la continuité et la variabilité du message parlé. C'est un élément de superforme, un atome phonatoire. Cependant, le phonatome ne peut être défini comme la transition entre deux phonèmes, car cette définition suppose implicitement la prééminence perceptive du phonème. Le phonatome peut être perçu alors même que les deux phonèmes-frontières n'ont pas d'existence acoustique. Par exemple, le mot "OUI" /wi/ possède un squelette sémantique schématisé sur la figure IV-18a, que l'on peut synthétiser facilement (voir chapitre V). La juxtaposition des deux phonèmes /w/ et /i/ (fig. IV-18b) ne conduit pas à la perception du mot /wi/, mais aux deux sons /w/ et /i/ séparés par un clic. Par contre, la seule partie intermédiaire, qui est le coeur du phonatome, suffit à faire percevoir le mot /wi/ (fig. IV 18c). Encore avons-nous considéré dans cet exemple des phonèmes quasi-stationnaires (voyelles). Pour des phonèmes transitoires (plosives, nasales) la démonstration est évidente. D'ailleurs, un phonème transitoire, /t/ par exemple, n'a pas d'existence autonome, et s'appuie toujours sur un phonème quasi-stationnaire. Pour faire comprendre à un enfant les caractères du son /t/, en vue de l'écriture, on lui présente en réalité une série de phonatomes /ta/, /tœ/, /ti/, /to/, /ty/. Le phonème-abstraction /t/ est défini en fait par l'ensemble de ses combinaisons avec les autres sons de la langue.

Il faut encore remarquer que les maxima de stabilité du squelette sémantique ne correspondent pas exactement aux phonèmes. Dans /aty/ par exemple (fig. IV-17), les maxima se situent avant et après l'explosion proprement dite, alors que le phonème /t/ est défini classiquement par son explosion.

Nous utiliserons donc la notation phonétique pour repérer les phonatomes

..../

et les messages phonétiques, sans que cela implique une quelconque correspondance entre les symboles phonétiques et les éléments de parole.

IV.3.2.5. Anamorphoses du squelette sémantique (bib. III-8, IV-12)

Le squelette sémantique est une structure temps-fréquence; la dimension de niveau y est réduite à sa plus simple expression. Les déformations, ou anamorphoses, peuvent donc affecter les dimensions de fréquence et de temps. Nous distinguerons quatre types d'anamorphoses, schématisées fig. IV-19, selon la dimension qu'elles touchent et leur régularité.

a) Anamorphose fréquentielle régulière (fig. IV 19b)

Cette déformation se ramène à un simple changement d'échelle en fréquence, ou affinité. Elle peut être définie par l'expression

$$X_2 \ (t,F) = X_1 \ (t, kF)$$

dans laquelle $X_1 \ (t,F)$ et $X_2 \ (t,F)$ représentent les squelettes sémartiques avant et après transformation, k étant une constante positive voisine de l'unité. On peut observer cette anamorphose dans les voix réelles : entre une voix d'homme et une voix d'enfant, il existe une affinité dont le rapport k peut atteindre 2. Ce phénomène peut également être étudié par synthèse : une expérience décrite en détail au chapitre V montre que l'intelligibilité de la voix de l'Icophone reste sensiblement invariante dans des affinités de 0,7 à 1,4, soit dans un rapport de 2 entre extrêmes (bib.IV-13).

b) Anamorphose fréquentielle irrégulière (fig. IV 19c)

Dans cette catégorie entrent toutes les transformations du type

$$X_2 \ (t,F) = X_1 \ \left[t, \ u \ (F)\right]$$

La fonction $u \ (F)$ définit la transformation effectuée sur la dimension de fréquence. C'est une fonction monotone croissante, qui dans la réalité ne s'écarte jamais beaucoup d'une droite passant par l'origine. En d'autres termes, l'anamorphose régulière constitue en général une première approximation de cette transformation, et l'on peut poser

$$u \ (F) = kF + \varepsilon \ (F)$$

Le terme $\varepsilon \ (F)$ représente l'écart par rapport à l'anamorphose régulière.

Nous appliquerons une telle transformation au chapitre VI, pour augmenter la définition de l'analyse dans la zone sensible.

Dans le cas des voyelles étudiées au § IV-22 d, nous avons déterminé les anamorphoses fréquentielles optima de la manière suivante : si F est la fréquence d'un formant du locuteur-référence, on porte en $u \ (F)$ la fréquence du même formant, dans la même voyelle, pour le locuteur donné (figure IV-20 a et b). En réduisant les courbes de régression à des droites, on trouve que l'échelle de fréquence du locuteur 2 doit subir la transformation

...../

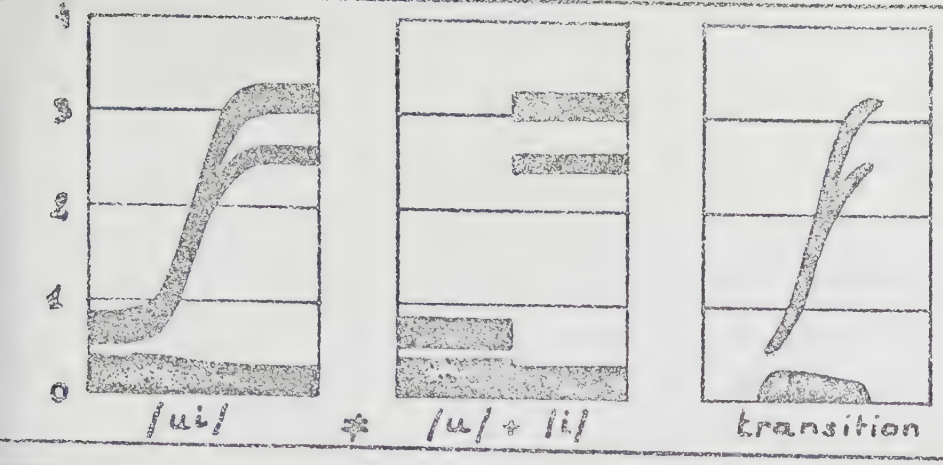


Fig IV-18

Squelette sémantique du mot "oui" schématisé. Ce mot ne peut être reconstitué à partir des seuls phonèmes /u/ et /i/ juxtaposés. Par contre la transition suffit à une synthèse intelligible.

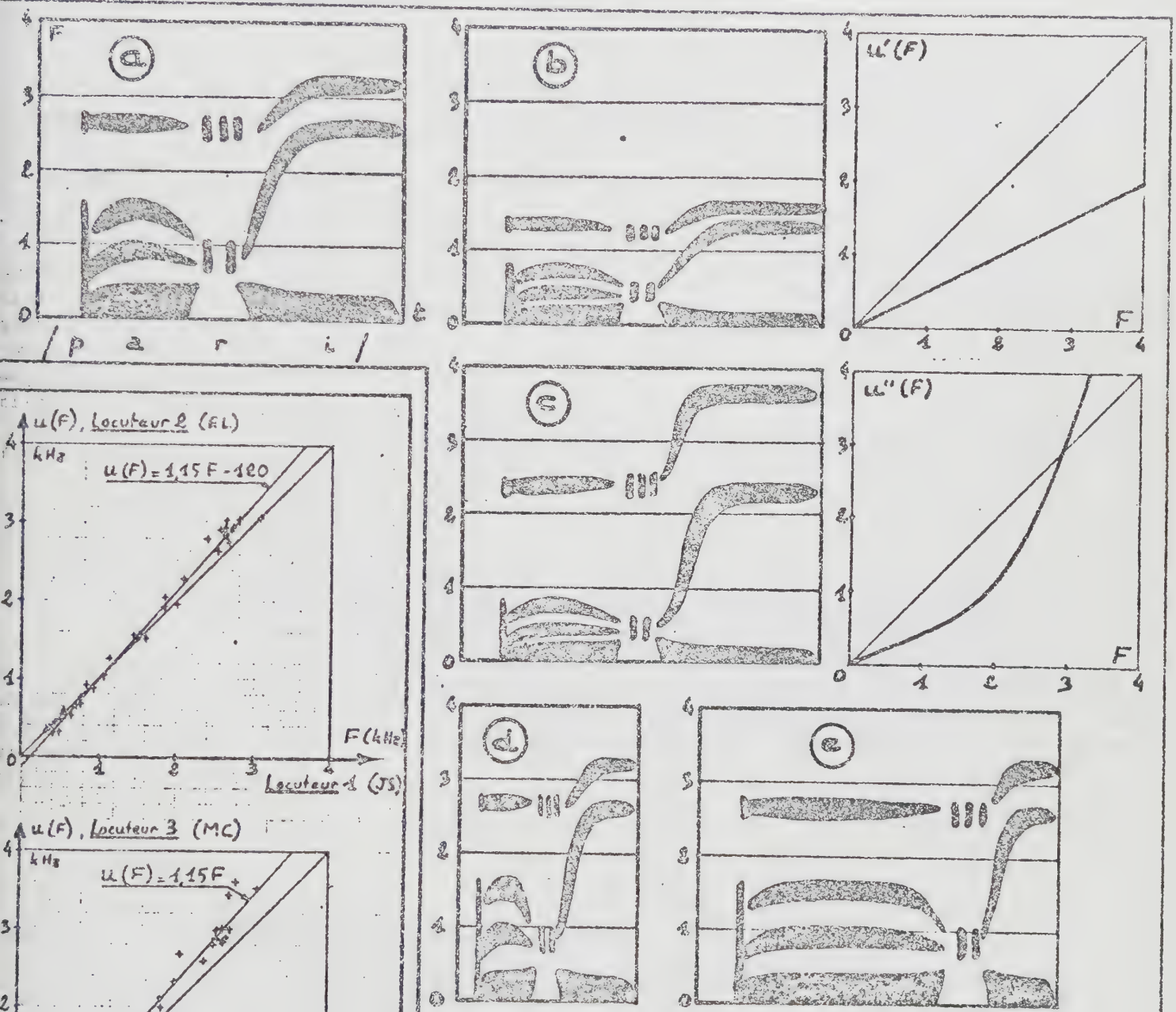
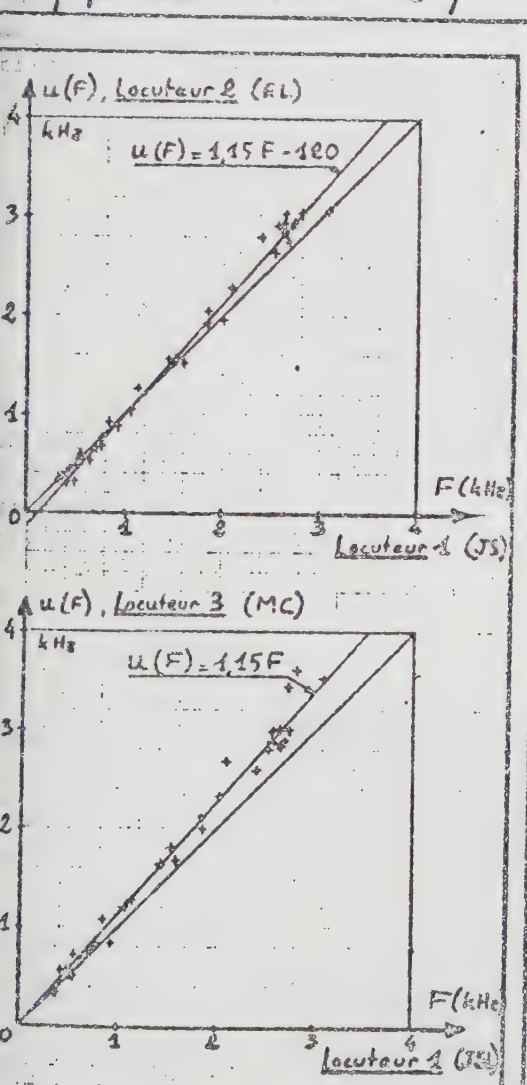


Fig IV-19

Schématisation des quatre types d'anamorphoses. a) Squelette sémantique du mot "Paris". b) Anamorphose fréquentielle régulière, de coefficient 0,5 ,accompagnée de la courbe u'(F) qui représente la transformation. c) Anamorphose irrégulière, et courbe u"(F) : les proportions ne sont plus conservées. d) Anamorphose temporelle régulière. e) Anamorphose temporelle irrégulière (anamorphose rythmique).



g IV-20

Etablissement des anaorphoses fréquentielles ilisées fig IV-11.

$$u_2 (F) = 1{,}15\ F - 120 \quad \text{(fréquences en Hz)}$$

Cette transformation, quoique linéaire, n'est pas une affinité. Elle indique que le 1er formant du locuteur 2 est généralement plus grave que celui du locuteur 1, et que les autres sont plus aigus, ce qui provient vraisemblablement des proportions anatomiques propres à chacun. Par contre, le locuteur n° 3 (féminin) donne lieu à une anamorphose définie par $u_3 (F) = 1{,}15F$ qui se réduit à une affinité si on laisse de côté l'accentuation fréquentielle du 3e formant dans les nasales /ã/ et /õ/, accentuation caractéristique de ce locuteur.

Si minimes qu'elles soient, ces opérations montrent la validité de la notion d'anamorphose irrégulière, même appliquée à des locuteurs voisins. On trouverait entre des voix très différentes (homme âgé et enfant) des résultats de même nature, mais plus spectaculaires.

Remarque I :

Si l'on n'impose pas à la fonction $u (F)$ d'être monotone et croissante, celle-ci ne définit plus véritablement une anamorphose du squelette sémantique, mais s'apparente à un transcodage. Par exemple, la fonction

$$u_1 (F) = F_1 - k_1 \ F$$

a pour effet dans sa partie positive le retournement de l'axe des fréquences : le grave devient aigu, et réciproquement.

Remarque II :

L'anamorphose fréquentielle n'est pas fixée une fois pour toutes pour un locuteur donné. La dispersion assez importante observée précédemment sur les fréquences des formants vocaliques (§ IV-22 d) indique au contraire que la transformation fréquentielle peut évoluer au cours du temps. Ce phénomène correspond à l'"intonation formantique" observée en voix chuchotée (bib. IV-16) et peut être mis en évidence par synthèse (bib. III-14, IV-17). Il serait donc raisonnable de noter l'anamorphose fréquentielle la plus générale par

$$X_2 (t,F) = X_1 \left[t, u (F,t) \right]$$

c) Anamorphose temporelle régulière (fig. IV-19d)

Il s'agit là d'une simple dilatation de l'axe des temps que l'on peut noter

$$X_2 (t,F) = X_1 \quad (rt,F)$$

Le facteur r représente la durée relative de X_2 par rapport à X_1. Si le message de référence X_1 est prononcé à raison de 10 symboles phonétiques par seconde, la valeur $r = 0{,}5$ donnera une élocution très rapide, et $r = 5$ donnera une élocution très lente. Le rapport entre extrêmes est donc de l'ordre de 10, ce qui permet un champ de liberté considérable.

...../

d) Anamorphose temporelle irrégulière, ou anamorphose rythmique (fig. IV-19c)

Comme précédemment pour la fréquence, nous pouvons définir une fonction de transformation temporelle $v(t)$, monotone et croissante, qui produit l'anamorphose

$$X_2 (t,F) = X_1 [v(t), F]$$

On peut mettre en évidence la durée relative r :

$$v(t) = rt + \lambda(t)$$

Mais ici le terme non linéaire $\lambda(t)$ ne peut que très rarement être considéré comme petit. En effet, selon le rythme adopté par le locuteur pour des besoins expressifs, la durée de certains sons du message est susceptible de grandes variations : celles-ci peuvent être de l'ordre de 10 pour les voyelles, mais sont nécessairement plus faibles pour les sons transitoires, en particulier pour les plosives. Ce problème est essentiel en reconnaissance automatique (normalisation temporelle).

IV.3.2.6. Champs de liberté de la voix humaine

Le recensement de toutes les anamorphoses possibles du squelette

$$X_2 (t,F) = X_1 [v(t), u(F,t)]$$

nous fournit déjà plusieurs champs de liberté que le locuteur humain ne se fait pas faute d'utiliser.

Mais, pour arriver à la notion de squelette sémantique, nous avons dû laisser de côté le spectre de raies, le niveau et le timbre de la voix. Chacun de ces paramètres évolue autour d'une valeur moyenne, dans divers champs de liberté qui permettent de transmettre des informations diagnostiques et esthétiques. Le fondamental Fo du spectre de raies, par exemple, oscille autour d'une valeur moyenne $\overline{Fo}$ qui constitue une caractéristique physique du locuteur, alors que la fonction $\frac{Fo\ (t)}{\overline{Fo}}$ permet de transmettre un autre type d'information, l'intonation, pratiquement indépendante de l'information phonétique.

Il n'est pas dans notre propos de décrire exhaustivement toutes les possibilités d'expression présentes dans le message parlé. Nous voulons seulement, par quelques exemples, en illustrer la complexité structurale, de façon à bien situer les problèmes de synthèse et de reconnaissance automatique. Ces exemples sont portés sur la fig. IV-21 et reprennent le même matériau phonétique que la figure IV-16 (début de la phrase "As-tu vu ce fameux lapin ?").

...../

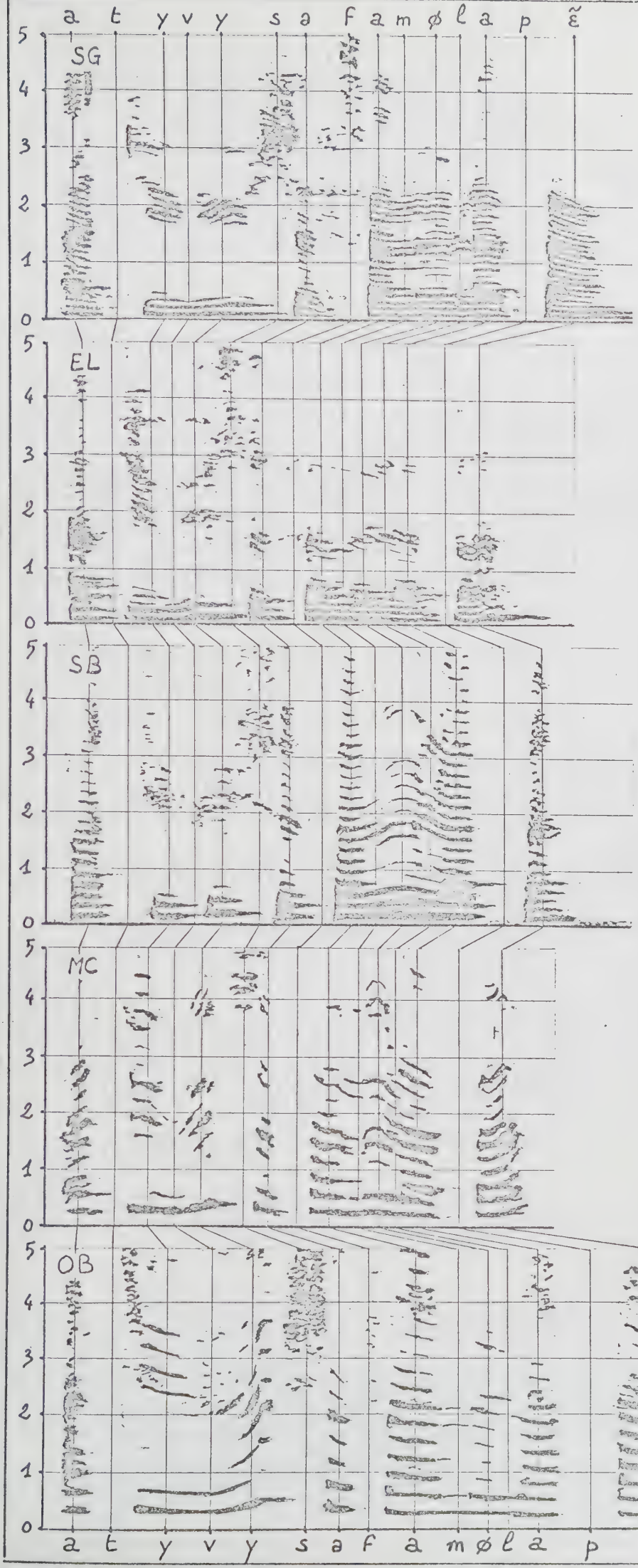


Fig IV-21

La même phrase ("As-tu vu ce fameux lapin"), prononcée par différents locuteurs. De haut en bas : deux voix masculines, deux voix féminines, une voix d'enfant (5 ans). Les frontières des phonatomes sont indiquées par des traits verticaux. On peut observer sur ces documents les divers types d'anamorphoses signalés dans le texte. On pourra se reporter à la figure IV-16 pour d'autres réalisations de la même phrase, et à la figure IV-17 pour son squelette sémantique.

Echelles:

- ordonnées : en kHz
- abscisses : en secondes

0 0,5 1s

	A	E	I	+	O	U	W	/	*	1	)	(	M	N	R	L	Y	X	S	J	7	F	V	P	B	T	D	K	G	Ç	Total
A	2	2	0	1	0	1	1	0	1	1	6	4	35	23	123	76	15	5	80	37	14	24	44	48	21	58	41	47	5	3	732
E	4	109	2	1	4	0	0	0	0	4	0	0	81	44	78	96	0	8	98	10	7	18	37	77	19	42	100	68	9	0	928
I	18	31	17	7	4	2	0	42	12	11	23	25	18	30	33	75	8	4	64	6	25	14	18	6	11	45	23	34	2	24	648
+	2	0	0	0	0	0	0	0	0	0	1	0	2	1	7	1	0	2	10	0	2	2	4	5	2	13	11	4	0	0	77
O	0	0	1	0	0	0	0	0	0	0	0	0	28	23	56	20	0	3	11	2	8	1	1	7	11	15	11	9	1	0	218
U	1	0	38	0	1	0	0	0	2	1	4	1	6	34	44	14	0	0	17	3	7	5	3	9	5	10	11	11	2	8	247
W	93	1	5	1	0	0	0	0	0	5	0	1	1	0	54	10	0	0	10	4	0	1	21	2	5	14	5	3	0	0	253
/	1	0	0	0	1	0	0	0	1	0	3	2	1	15	0	10	0	2	20	1	4	8	4	18	7	22	24	10	1	0	163
*	0	0	0	0	0	0	0	0	0	1	4	0	4	11	2	33	0	7	50	8	5	10	7	18	11	64	39	19	4	0	306
1	2	1	0	0	0	0	0	0	0	0	0	0	2	11	2	1	0	1	17	4	2	6	0	17	0	17	14	5	1	0	114
)	12	2	0	3	0	0	0	0	2	1	59	2	20	9	46	33	0	4	25	8	24	11	8	31	6	47	36	29	8	24	464
(	1	0	0	0	1	0	0	0	0	0	1	1	20	18	94	86	13	0	49	2	17	8	5	11	2	69	27	31	7	1	474
M	37	44	27	1	13	10	15	11	40	1	15	52	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	296
N	31	63	32	1	10	18	11	7	10	2	22	24	0	17	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	8	266
R	66	110	76	4	27	13	28	9	40	18	52	67	15	9	20	27	0	2	41	7	5	5	8	10	3	41	39	13	3	27	799
L	133	133	52	6	18	31	11	7	21	8	74	44	5	3	2	66	0	1	5	1	2	3	4	3	1	6	4	7	2	18	685
Y	2	4	2	0	2	0	0	0	1	1	6	3	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	10	40
X	10	7	8	0	1	0	0	0	4	0	10	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	50
S	35	101	110	6	17	35	22	31	41	5	37	56	0	0	0	0	0	0	6	0	0	1	0	19	0	37	0	14	0	21	614
J	12	41	6	2	0	4	15	1	7	0	12	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	117
Z	26	10	28	8	5	14	2	7	17	10	13	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	159
F	16	2	31	2	11	5	5	3	7	6	4	20	0	0	8	5	0	0	1	0	0	13	0	0	0	0	0	0	0	0	150
V	12	15	34	2	6	2	31	0	15	4	7	40	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	184
P	81	20	17	1	17	11	36	2	12	1	11	19	0	0	65	33	0	0	3	0	0	0	0	12	0	4	0	0	0	0	350
B	8	1	13	5	8	7	10	5	2	0	2	4	0	0	13	25	0	0	6	0	0	0	0	0	0	3	0	0	0	0	119
T	56	55	55	11	13	27	31	4	39	10	54	55	2	0	102	0	0	0	0	0	8	0	0	0	0	24	0	0	0	17	573
D	23	186	44	1	11	43	4	8	37	13	54	19	1	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	471
K	43	49	38	0	37	14	20	36	8	5	5	18	0	1	15	8	0	0	24	0	0	1	0	0	0	10	0	6	0	10	357
G	8	1	1	1	2	1	1	1	1	0	0	1	0	0	20	2	0	0	0	0	5	0	0	0	0	0	0	0	0	0	54
Ç	4	0	1	1	0	2	0	0	1	0	5	2	4	0	0	4	0	0	8	0	2	1	3	6	1	2	18	2	0	0	77
Total→	7541	1001	653	77	219	250	254	182	332	119	497	485	275	257	820	634	40	44	555	99	159	142	175	308	112	552	414	323	52	202	10000

1	DE	166	POUR 10000
2	LE	133	POUR 10000
3	LA	133	POUR 10000
4	AR	123	POUR 10000
5	SI	119	POUR 10000
6	RE	110	POUR 10000
7	EE	109	POUR 10000 ←
8	TR	102	POUR 10000
9	SE	101	POUR 10000
10	ED	100	POUR 10000
11	ES	98	POUR 10000
12	EL	96	POUR 10000
13	(R	94	POUR 10000
14	WA	93	POUR 10000
15	(L	86	POUR 10000
16	PA	81	POUR 10000
17	EM	81	POUR 10000
18	AS	80	POUR 10000
19	ER	78	POUR 10000
20	EP	77	POUR 10000
	RI	76	POUR 10000
	AL	76	POUR 10000
	IL	75	POUR 10000
	LJ	74	POUR 10000
25	(T	69	POUR 10000
	EK	68	POUR 10000
	RI	67	POUR 10000
	LL	66	POUR 10000 ←
	RA	66	POUR 10000
30	PR	65	POUR 10000
	*T	64	POUR 10000
	IS	64	POUR 10000
	NE	63	POUR 10000
	))	59	POUR 10000 ←
35	AT	58	POUR 10000
	TA	56	POUR 10000
	S(	56	POUR 10000
	OR	56	POUR 10000
	T(	55	POUR 10000
40	TI	55	POUR 10000
	TE	55	POUR 10000
	D)	54	POUR 10000
	T)	54	POUR 10000
	WR	54	POUR 10000
45	LI	52	POUR 10000
	R)	52	POUR 10000
	M(	52	POUR 10000
	*S	50	POUR 10000
	KE	49	POUR 10000
50	(S	49	POUR 10000
	AP	48	POUR 10000
	IT	47	POUR 10000
	AK	47	POUR 10000
	JR	46	POUR 10000
55	IT	45	POUR 10000
	DI	44	POUR 10000
	L(	44	POUR 10000
	ME	44	POUR 10000
	UR	44	POUR 10000
60	EN	44	POUR 10000

Fig IV-22

Fréquence d'occurrence des phonatomes de la langue française (ci-dessus) et classement par fréquence d'occurrence décroissante (ci-contre).

Ces comptages ont été effectués à partir de divers textes français, après une transcription phonétique automatique faisant apparaître 16 913 phonatomes. La fréquence (en dix-millièmes) du phonatome "R)" (ou /re/ en notation internationale) se trouve à l'intersection de la ligne "R" et de la colonne ")". La fréquence anormalement élevée des phonatomes "EE","LL" et "))", signalés par des flèches, est imputable à une erreur du programme de traduction phonétique, mais n'influence que très peu l'ensemble des résultats.

IV.4. STATISTIQUE PHONETIQUE ET MOTS ALEATOIRES

Ayant mis en évidence la notion de phonatome sur le plan acoustique, nous allons maintenant étudier ses répercussions sur le plan linguistique. Il s'agit de montrer, d'une part que la distribution des digrammes phonétiques dans la langue française est soumise à certaines normes et, d'autre part, que la connaissance de ces normes donne lieu à des applications fructueuses dans divers domaines (synthèse, reconnaissance, création linguistique).

IV.4.1. Statistique phonétique

Nous adoptons dans ce paragraphe une démarche statistique, celle-là même qui est à la base de la théorie de l'Information. C'est dire, selon les réserves exprimées au chapitre II, qu'il n'est pas question ici de communication individuelle. Ce type de démarche permet à la linguistique quantitative de dégager certaines tendances inconscientes de la langue. Ainsi la loi de Zipf-Mandelbrot (bib. IV-17, II-12), qui établit la constance du produit de la fréquence d'occurrence par le rang de chaque mot (rang dans un classement par fréquence d'occurence décroissante), peut-elle être interprétée comme une tendance vers l'efficacité maxima des messages (principe de Zipf ou du moindre effort).

IV.4.1.1. Comptage des digrammes phonétiques

La répartition des phonèmes dans les textes ou dans le discours a été souvent étudiée. Il est bien connu que les phonèmes /r/, /ə/, /l/, /a/, /s/, sont parmi les plus fréquents, et /ʃ/, /œ/, /g/ parmi les plus rares.

Nous avons étudié la répartition des digrammes dans divers textes français (bib. IV-14, IV-18), et montré que les fréquences d'occurrence sont soumises à une norme, qui constitue une caractéristique de la langue. La stabilité de la répartition observée nous autorise à parler de probabilité lorsque le corpus utilisé est suffisamment long.

La figure IV-22 montre le tableau des fréquences d'occurrence ou probabilités de digrammes obtenu après un tel comptage. Le matériau phonétique provient de texte français variés transcrits phonétiquement par le programme de D. TEIL (cf chapitre V), comportant en tout 16.913 digrammes répartis en 4489 mots. On a adopté ici $N = 30$ phonèmes de base $/\varphi_n/$, y compris le /ə/ dit "muet" (symbole machine &), qui théoriquement devraient fournir $N^2 = 900$ digrammes différents.

La probabilité (en dix-millièmes) du digramme $/\varphi_i\ \varphi_j/$ est portée en p_{ij}, à l'intersection de la ligne i et de la colonne j - Les sommes par ligne

$$s_i = \sum_{j=1}^{N} p_{ij}$$ et par colonne $$\sigma_j = \sum_{i=1}^{N} p_{ij}$$ représentent

respectivement les probabilités de trouver le phonème $/\varphi_i/$ en première position et le phonème $/\varphi_j/$ en seconde position du digramme. Pour un même phonème $/\varphi_n/$

..../

on a généralement $s_n \neq \sigma_n$; l'égalité serait obtenue si l'on comptait comme un phonème particulier le caractère séparateur des phases (point ou virgule). L'ensemble du corpus étudié serait ainsi une suite ininterrompue de phonèmes, et chaque phonème apparaîtrait autant de fois en seconde qu'en première position, à deux unités près (début et fin). On pourrait ainsi se débarrasser de la notion de /ə/ "muet", peu satisfaisante acoustiquement. Pratiquement, nous ne ferons pas une grande erreur en prenant comme probabilité du phonème $/\varphi_n/$ le maximum de s_n ou σ_n

$$p_n = \max (s_n , \sigma_n)$$

La figure IV-22 comporte également un classement des digrammes par probabilité décroissante. La répartition des digrammes présente deux caractéristiques remarquables. La première est le nombre limité de phonatomes "utiles" : sur les 900 digrammes possibles nous n'en avons rencontré que 613 différents, dont 67 une seule fois. Des études analogues menées ultérieurement sur d'autres bases (J.P. HATON, bib. IV-19, sur 1096 digrammes possibles, et J.P. TUBACH, bib. IV-20, sur 1225 digrammes possibles, tous deux sur des corpus phonétiques plus importants que le nôtre, ont conduit également à ce chiffre de 600 phonatomes utiles.

La seconde caractéristique est la régularité de la distribution : la courbe $p_{ij} = f (\log r_{ij})$, où r_{ij} est le rang du digramme $/\varphi_i \ \varphi_j/$ dans un classement par probabilité décroissante, tend très vite vers une droite, sauf dans le cas de textes trop courts ou de "langages monstres" (fig. IV-23). Ces derniers sont des "jargons" ou langages techniques utilisant certains mots ou certains sons plus fréquemment que les autres. On rencontre d'ailleurs le même phénomène en poésie.

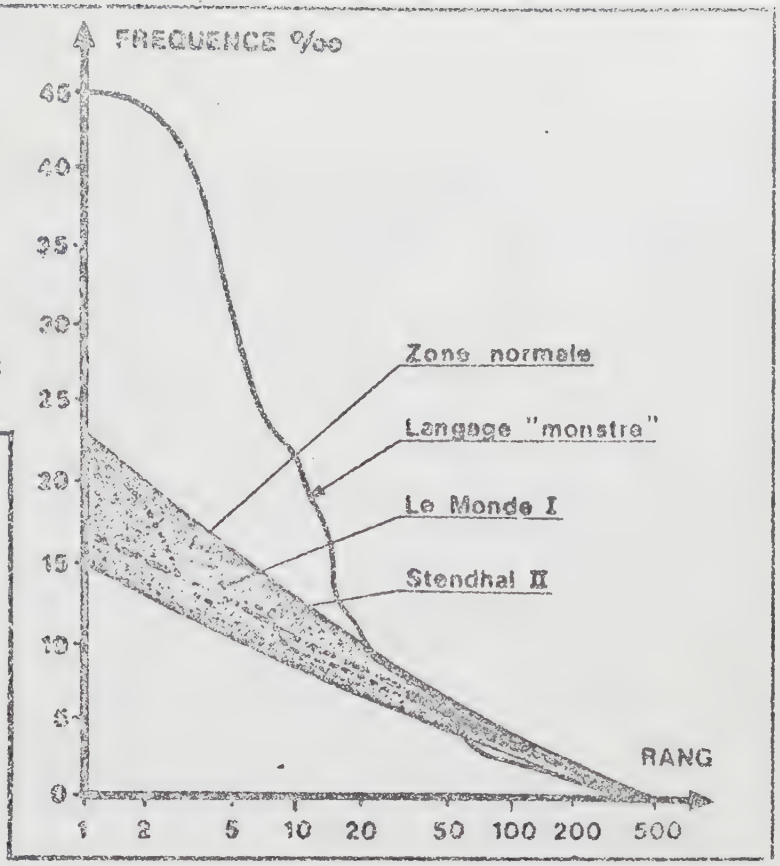
Les statistiques sur les trigrammes phonétiques sont plus difficiles à mettre en oeuvre et à interpréter car si l'on considère 30 phonèmes on obtient $30^3 = 27\,000$ trigrammes phonétiques théoriquement possibles. Ce travail a néanmoins été abordé par J.P. TUBACH, qui, sur un corpus suffisamment important (88 000 phonèmes), a mis en évidence les trigrammes les plus fréquents: /par/, /siɔ̃/, /dəl/, /əla/, /sɛt/, /trə/, /pur/, etc...

La précision que l'on peut obtenir en machine ne doit pas faire illusion quant à l'exactitude des résultats. En effet, la transcription phonétique est une opération très délicate, qui dépend de l'accent et de la culture du locuteur et du transcripteur. La vitesse d'élocution joue également un rôle important : certains phonèmes, certaines syllabes sont souvent presque élidés. En réalité, le transcripteur ne peut être parfaitement objectif, et il ne serait pas raisonnable d'accorder aux résultats une trop grande valeur quantitative, même si l'intervalle de confiance (de l'ordre de quelques millièmes, pour un corpus moyen) paraît satisfaisant.

La répartition des mots en fonction de leur longueur phonétique est plus précise et plus facile à déterminer. La figure IV-24 établie à partir de nos relevés montre que la langue écrite ou parlée emploie une nette majorité de mots de deux phonèmes, alors que le même relevé effectué dans un dictionnaire fait apparaître un maximum pour les mots de 4 phonèmes. Cette anomalie provient de l'utilisation des mots-chevilles, qui marquent les relations entre les mots-idées. Ces mots (articles, pronoms, prépositions, verbes courants, etc...) sont des mots courts, de un à trois phonèmes, que l'on répète très fréquemment. Ainsi s'explique la forme singulière de la courbe IV-24, qui laisse quand même apparaître un second maximum correspondant à la longueur modale des mots-idées. La même remarque s'applique à la liste des digrammes les plus fréquents (fig. IV-22) : parmi ceux-ci on rencontre de nombreux mots-chevilles ("de" ou "deux", "le", "la", "se", "ce" ou "ceux", "pas", "il", "elle" etc...)

...../

Fig IV-23 Fréquence d'occurrence en fonction du rang des phonatomes (bib IV-14). La quasi-totalité des textes étudiés se situent dans une zone restreinte (zone "normale"), dont l'existence est dûe aux contraintes phonétiques de la langue française. Certains textes ("langages-monstres", ini une page de physique) s'écartent notablement de la norme.



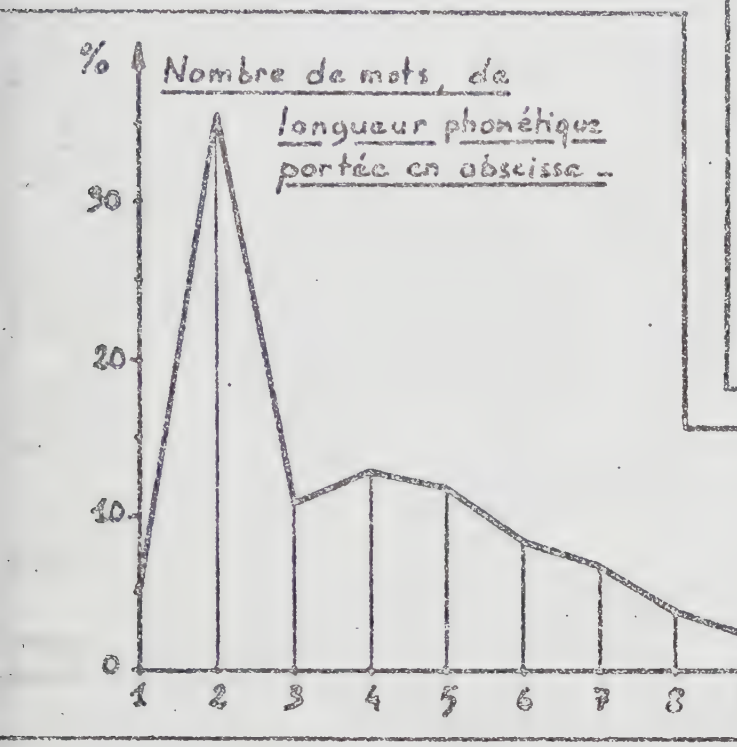
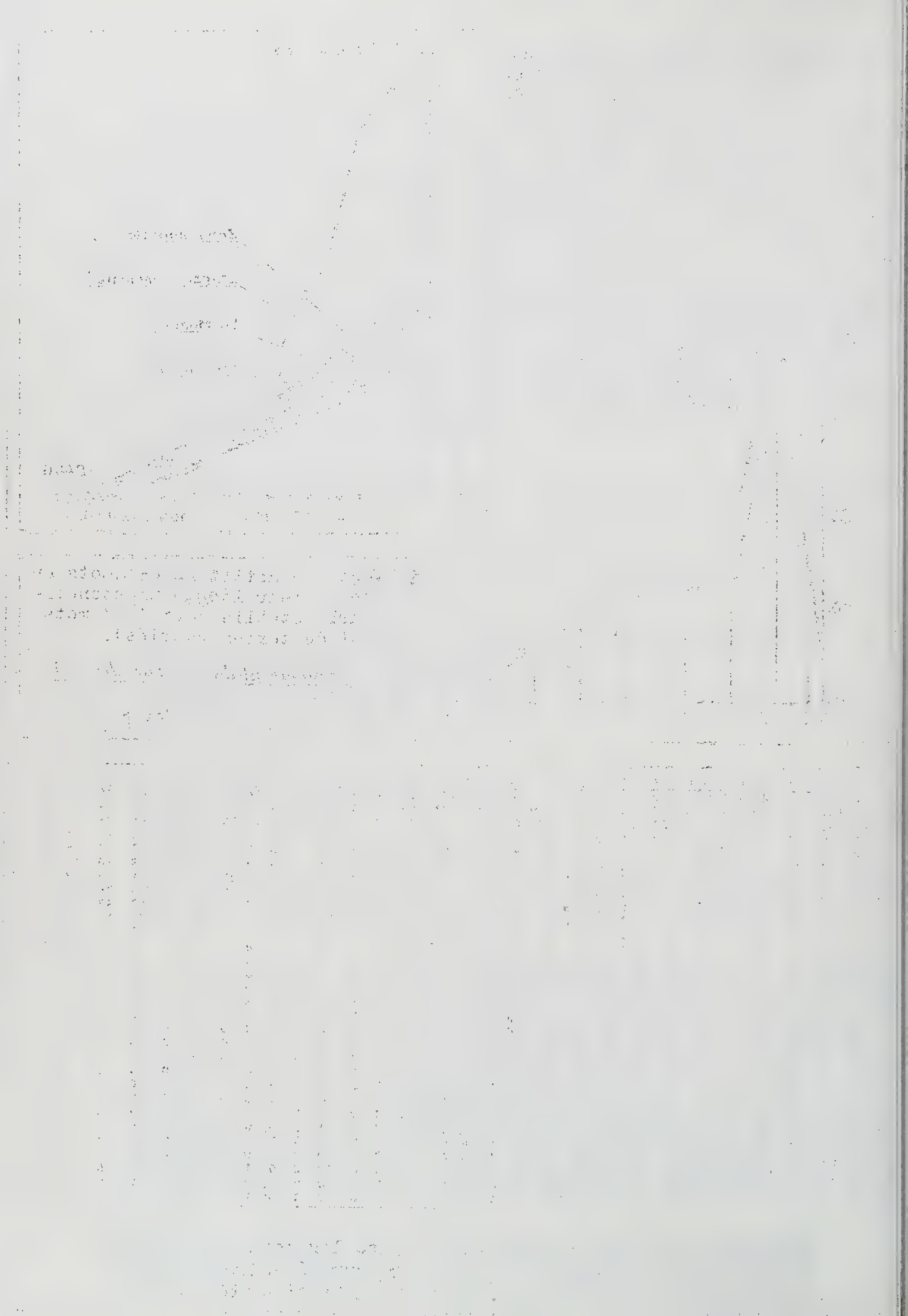


Fig IV-24 Répartition des mots en fonction de leur longueur phonétique (courbe établie sur 4563 mots provenant de textes variés).

A	E	I	+	O	U	H	/	@	1	b	ʃ	N	N	ɛ	ɛ̃	Y	X	S	J	Z	F	V	P	B	T	D	K	G	ʃ
2	1	0	2	1	0	3	0	1	1	6	3	35	23	123	75	17	4	83	40	9	24	43	48	20	61	38	47	6	4
3	102	4	1	5	0	0	0	0	5	0	0	76	45	81	104	0	10	106	6	6	16	41	75	24	43	109	74	10	0
21	39	18	6	4	3	0	39	14	10	20	29	20	30	21	82	9	1	60	5	28	13	19	3	11	45	26	33	2	26
2	0	0	0	0	0	0	0	1	0	1	0	3	2	6	2	0	2	12	0	1	2	4	4	2	14	12	4	0	0
0	0	1	0	0	0	0	0	0	0	0	0	31	26	59	17	0	4	11	1	6	0	2	5	10	13	13	6	2	0
0	0	36	0	1	0	0	0	2	0	3	0	0	44	44	16	0	0	16	4	6	4	4	8	4	9	10	10	2	11
98	0	13	0	0	0	0	0	0	3	0	1	1	9	47	10	0	0	9	4	8	3	25	2	5	14	5	2	0	0
1	0	0	0	1	0	0	0	1	0	4	2	1	17	0	12	0	1	20	0	5	11	5	18	8	26	28	10	2	0
0	0	1	0	0	0	0	0	0	1	3	0	4	12	3	35	0	7	57	8	4	0	8	23	13	67	42	21	5	0
3	2	0	0	0	0	0	0	0	1	0	0	3	7	3	0	0	1	18	4	2	7	0	14	0	21	14	8	1	0
14	3	0	3	1	0	0	0	3	2	79	3	20	8	54	39	0	2	28	8	24	12	7	33	5	53	41	31	9	26
1	0	0	0	1	0	0	0	0	0	1	2	20	18	106	87	16	0	50	1	15	6	6	11	3	70	26	31	8	0
36	45	24	0	12	12	17	11	37	2	15	54	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
30	74	28	1	8	16	13	8	11	2	22	25	0	10	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	10
79	126	76	7	30	12	32	13	51	22	56	76	12	6	18	20	0	0	30	6	4	3	6	5	3	29	27	9	1	19
136	136	51	5	24	30	11	8	15	9	77	37	6	1	1	50	0	0	3	1	2	3	3	2	1	2	1	5	2	11
4	4	2	0	2	0	2	0	1	1	7	3	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	9
8	5	6	0	0	0	0	0	5	0	5	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
30	104	113	8	15	30	20	30	44	5	39	90	0	0	0	0	0	0	3	0	0	0	0	18	0	25	0	15	0	16
8	40	7	2	0	2	12	1	7	0	10	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
23	8	25	8	3	13	4	11	10	8	13	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
14	4	28	3	9	6	4	3	9	8	5	19	0	0	6	4	0	0	1	0	0	10	0	0	0	0	0	0	0	0
14	14	36	2	7	3	32	0	13	4	8	35	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
81	18	15	1	15	8	33	2	11	0	11	21	0	0	55	24	0	0	2	0	0	0	0	9	0	3	0	0	0	0
8	1	11	6	7	7	12	11	2	0	1	2	0	0	14	21	0	0	5	0	0	0	0	0	0	4	0	0	0	0
60	49	53	12	0	31	30	7	43	12	34	60	2	0	62	0	0	0	0	0	7	0	0	1	0	16	0	0	0	12
20	184	39	0	11	43	2	0	35	11	55	29	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
44	45	41	0	34	14	18	32	8	4	5	20	0	0	12	6	0	0	20	0	0	1	0	0	0	7	0	6	0	3
9	0	1	4	4	1	2	1	2	0	0	0	0	0	16	3	0	0	0	0	4	0	0	0	0	0	0	0	0	0
22	2	6	4	4	13	0	0	11	0	30	9	3	0	0	2	0	0	8	0	0	2	3	5	0	2	16	4	0	0

Fig IV-25 Comptage des phonatomes dans les mots aléatoires, permettant, par comparaison avec la figure IV-22, de vérifier le bon fonctionnement du programme de génération aléatoire.



IV.4.1.2. Contraintes phonétiques

Les relevés précédents montrent à l'évidence que les phonèmes ne sont pas équiprobables, et qu'ils ne sont pas statistiquement indépendants. Si les phonèmes $/\varphi_n/$ et $/\varphi_m/$, de probabilité p_n et p_m, étaient indépendants, la probabilité des digrammes $/\varphi_n \varphi_m/$ et $/\varphi_m \varphi_n/$ serait

$$P_{mn} = P_{nm} = p_m \cdot p_n$$

On vérifie aisément sur les relevés que ces relations ne sont généralement pas vérifiées. Considérons par exemple les phonèmes /t/ et /r/, et les digrammes /tr/ et /rt/, dont les probabilités sont les suivantes :

$p(t) = 0{,}056$ $\qquad$ $p(r) = 0{,}081$

$p(tr) = 0{,}010$ $\qquad$ $p(rt) = 0{,}004$

Ces deux dernières valeurs peuvent être rapportées au produit $p(t).\ p(r)$, qui représente la probabilité de /rt/ et /tr/ si aucune contrainte n'existait dans l'association des phonèmes :

$$\frac{p(tr)}{p(t).p(r)} = 2{,}2 \qquad \frac{p(rt)}{p(t).p(r)} = 0{,}9$$

On voit ainsi que le digramme /tr/ est deux fois plus fréquent que l'on ne pourrait s'y attendre au simple vu des probabilités de phonèmes.

Des calculs analogues mènent à

$$\frac{p(d\,ə)}{p(d).p(ə)} = 17{,}6 \qquad \frac{p(ə\,d)}{p(ə).p(d)} = 0{,}12 \quad \text{etc....}$$

Ces écarts, qui peuvent être importants, reflètent certaines contraintes phonétiques de la langue. Il serait intéressant de chercher à quoi correspondent ces contraintes sur le plan acoustique. Une telle étude devrait tenir compte de la forme particulière de chaque phonatome, de la quantité d'information auditive qu'il transporte, de l'anatomo-physiologie du conduit vocal, et des habitudes articulatoires propres à chaque langue. Il est bien évident d'ailleurs que, si chaque phonatome présente un certain "coût" articulatoire - cette notion restant à préciser - les phonatomes les plus faciles à prononcer sont les plus fréquents, mais transportent le moins d'information. Un certain équilibre doit donc s'établir entre la tendance à l'économie et la tendance à la précision, équilibre que la régularité des courbes rang-fréquence d'occurrence permet de pressentir, et dont les paramètres méritent d'être déterminés.

IV.4.1.3. Utilisation des probabilités de digrammes

a) Synthèse

Nous verrons (chapitre V) que les fréquences d'occurrence ont été mises à profit dans l'Icophone numérique, pour optimiser la recherche des phonatomes

...../

en mémoire.

b) Reconnaissance

Lorsqu'un système de reconnaissance "hésite" entre deux phonèmes, phonatomes ou mots, on peut lui imposer de choisir le plus fréquent, ce qui améliore statistiquement les résultats. C'est une manière d'utiliser les contraintes phonétiques (ou sémantiques s'il s'agit de mots) de la langue. Cependant, cette ressource ne doit intervenir qu'en dernier lieu, car son origine est statistique : elle permet de rectifier surtout les messages formés d'éléments fréquents, donc porteurs de moins d'information, et risque de détruire les messages originaux. Son efficacité risque d'être illusoire, si l'on ne prend pas certaines précautions.

c) Tests phonétiques

Dans de nombreux cas (tests d'intelligibilité en pathologie, rééducation, télétransmission, acoustique des salles, etc...) on a besoin de listes de mots représentant correctement une langue sur le plan phonétique. Lorsque ces mots font partie de la langue, il convient de vérifier que la répartition des phonatomes est voisine de la norme. Mais on peut aussi créer de nouveaux mots, sans signification a priori, qui respectent cette norme par construction. Nous avons appelé "mots aléatoires" ces êtres phonétiques particuliers.

IV.4.2. Mots aléatoires

IV.4.2.1. Principe

Les mots aléatoires sont fabriqués en juxtaposant des digrammes phonétiques choisis selon leur probabilité d'occurrence dans la langue. Le programme écrit par D. TEIL, puis amélioré par H. LUCOT, choisit le phonème (φ_m) à partir du phonème précédent $/\varphi_n/$ sur le principe suivant :

La probabilité pour que $/\varphi_n/$ apparaisse en première position est s_n ; si p_{nj} est la probabilité du digramme $/\varphi_n\ \varphi_j/$, le rapport $\frac{p_{nj}}{s_n}$ représente la probabilité que $/\varphi_j/$ apparaisse en seconde position quand $/\varphi_n/$ est réalisé. Ce nombre est compris entre 0 et 1 .

A chaque phonème $/\varphi_j/$ nous allons associer un nombre B_j tel que

$$B_j - B_{j-1} = \frac{p_{nj}}{s_n}$$

Les nombres B_j sont donc définis par

$$\begin{cases} B_o = 0 \\ B_j = \frac{1}{s_n} \sum_{1}^{d} p_{nj} \end{cases}$$

...../

On remarquera que $B_n = 1$

La donnée d'un nombre aléatoire A compris entre 0 et 1 permet alors de choisir le second phonème $/\varphi_m/$: l'indice m est l'indice du nombre B_j immédiatement supérieur à A

$$B_{m-1} < A < B_m$$

L'ambiguïté qui pourrait provenir des valeurs $A = B_j$ est facilement levée en programmation : il suffit de multiplier toutes les valeurs (A et B_j) par un grand nombre (10^4 par exemple, puisque les probabilités sont données en dix-millièmes), puis de choisir B_j parmi les nombres entiers et A parmi les nombres irrationnels.

Le processus doit être amorcé : l'opérateur fournit le phonème de départ, la longueur et le nombre des mots aléatoires désirés, et le tableau des probabilités. La génération des nombres A est en fait pseudo-aléatoire : un second passage du programme avec les mêmes conditions initiales donne la même liste de mots.

Les mots obtenus présentent un grand intérêt pour l'établissement de tests phonétiques, ou pour la création de mots nouveaux auxquels on demande simplement de "sonner français". Voici quelques échantillons, retranscrits dans une orthographe arbitraire (nous avons éliminé ici les mots "accidentellement" français).

- mots de 3 phonèmes : déa, nance, laule, vran, nure
- mots de 4 phonèmes : anure, coufan, reté, vobi, boil, zizan, lara, luzan, fouaque
- mots de 5 phonèmes : larare, bousque, omance, massire, vesti, filure, aillard, fonelle, nanchor, tantoit, supil, laquefan etc...

Une seconde partie du programme effectue sur les listes de mots aléatoires les mêmes comptages que ceux effectués sur le corpus phonétique. On peut ainsi vérifier que le processus décrit plus haut est correct : la figure IV-25 donne le tableau des probabilités de digrammes, p'_{ij}, relevé après création de 10.000 mots aléatoires de longueur comprise entre 1 et 10 phonèmes, le tout représentant 45.000 digrammes. On comparera utilement ce tableau p'_{ij} au tableau p_{ij} utilisé comme donnée (fig. IV.22).

IV.4.2.2. Comptage des digrammes différenciés selon leur position dans le mot

E. LEIPP nous a suggéré que le digramme initial d'un mot français n'observait peut-être pas la même distribution que les digrammes suivants. Il est en effet vraisemblable que les digrammes initial et final, qui doivent assurer le raccordement au plus grand nombre possible de mots différents, n'obéissent pas aux mêmes contraintes phonétiques que les digrammes intérieurs au mot, assemblés en une forme unique. Nous avons donc entrepris de nouveaux comptages, en constituant sept tableaux de probabilité différents :

...../

- le tableau des digrammes initiaux (INIT)
- le " " " médians (MED)
- le " " " finals (IFIN)
- le " " " de liaison (LIAIS)
- le " " mots de un phonème (IP1)
- le " " " deux phonèmes (IP2)
- le " total (digrammes indifférenciés)

Après le comptage, le contenu du tableau IP2 est réparti pour moitié dans les tableaux INIT et IFIN.

Le résultat de ces comptages, effectués sur le même corpus phonétique que précédemment (16.913 digrammes) montre que les digrammes initiaux et finals ainsi que les mots de 2 phonèmes, sont en majorité du type consonne-voyelle, les digrammes de liaison du type voyelle-consonne, et les digrammes médians se répartissent entre voyelle-consonne, et secondairement consonne-voyelle.

Nous avons obtenu deux fois dix séries de mille mots aléatoires de longueur 1 à 10 phonèmes à partir du tableau total, puis à partir des tableaux INIT, MED et IFIN. La figure IV-26 permet de comparer les mots de 4 phonèmes obtenus par ces deux processus, toutes choses égales d'ailleurs.

Arrivés à ce point, nous avons senti la nécessité d'un critère objectif permettant d'estimer la "qualité" des mots aléatoires. Pour cela, nous avons compté dans chaque liste le nombre de mots aléatoires coïncidant avec des mots de la langue française (ces mots sont soulignés sur la figure IV-26). La figure IV-27 présente le résultat de ces relevés effectués par Mme A.M. LIENARD sur les listes précédentes. Si ce critère est valable, la différenciation des tableaux de probabilité constitue un progrès important dans la genèse des mots aléatoires.

Les taux portés sur la fig. IV-27 représentent les occurrences de mots français, rapportées au nombre de mots aléatoires créés dans chaque série. En fait, il se produit de nombreuses répétitions dans les listes de mots courts (de 1 à 3 phonèmes), et peut-être serait-il judicieux de les éviter, en vérifiant pour chaque mot aléatoire qu'il n'a pas été déjà créé dans la même liste. Cette amélioration ne changerait guère, semble-t-il, le sens de l'évolution marquée sur la fig. IV-27, car les taux de répétition deviennent négligeables dès que la longueur dépasse 3 phonèmes. Ce dernier point s'explique si l'on remarque que le nombre total de mots imaginables avec n symboles choisis dans un répertoire de N phonèmes est, en supposant tous les digrammes équiprobables :

Arrangements de N symboles n à n : $A_N^n = \frac{N\ !}{(N-n)\ !}$

Soit pour N = 30 , n = 1 ; $A_{30}^1 = 30$

n = 2 ; $A_{30}^2 = 30 \times 29 = 870$

n = 3 ; $A_{30}^3 = 30 \times 29 \times 28 = 24\ 360$

n = 4 ; $A_{30}^4 = 30 \times 29 \times 28 \times 27 = 657\ 720$

...../

SARI	AY)S	MORT	FLUL	OLLE	JERZ	(IF(	(PLO	O(IK	DATR
L)FR	NWKA	&VAN	NES(	OTUJ	AMAR	SOMA	)*J)	)LEL	*PAS
SAMM	NARD	WS*T	TAVI	KELI	AMIT	L(PO	T/DI	MMAS	RBIS
SOMM	EESW	MOI(	LIVI	J&UL	V)EM	*PAS	NNUF	NNWS	LURI
OLM(	S)TT	ALEE	LATE	RRIN	DMED	FISI	SAKE	N)LU	NA)P
WRSE	I*D)	IKS*	PWRR	VAYI	WAJ*	S*RA	LAVU	ZIKE	UR(L
T)))	LIMA	KS(L	SABI	PARS	APRE	ATAS	LATR	Z(RA	AKAD
(KR(	S*I*	AP*D	FADA	RWVI	PANA	RJEE	R*KU	ASI&	UNET
WJ&)	FI&S	M*BR	USIT	T)RE	F(ON	UII)	(GR*	LERD	M&AK
/SIT	ALIT	WARE	ARAR	KAL/	SIUN	)ALL	LUD)	OLIS	WIRA
MWVO	PAM(	R*L(	REGA	IS(L	RESW	OLFI	FKED	APOD	RRIX
)M/P	BSWL	T))R	REDE	)RAS	T(LE	IJEL	IZIT	*BAK	WARA
LWVO	LAK(	VWAY	EMEV	)M(I	)G/P	I(I*	+VAZ	KONI	OLLE
TEDM	+BOM	URU&	(LAT	ZORA	ENWA	)(L)	SFAT	NEM*	DIN)
TZAL	ES+D	PUII	S+SE	F*NI	M*DR	)ZI)	WBSE	(KOR	IIAL
(LWR	RWAJ	WAP)	KIZ)	S*K/	-*TR(	WPAJ	TR)D	AV)F	KARL
ELAK	/TRW	PR(P	AR/G	(SOT	PWIT	PR(K	M&(L	MI/K	RISE
ILLA	EST(	IB+K	UTEM	RTIZ	)PRE	Z)RE	I/P(	RELM	BSER
IRLA	LJ(L	ASI/	SELL	S(LA	V(DE	EKWL	AREN	SESA	KINE
ARES	M(S&	WALE	IT*S	ILUR	KUN+	S*T/	I(KE	ROLA	AJWV
1S)G	*KAZ	AR*R	ARWS	*DES	DI&K	)DEF	ERVI	LISA	K/ZI
AVWB	+DER	PWTW	UN(L	SKEE	Z*NE	*M)R	STIL	TINF	KILI
TAL+	+TWA	AK/G	RD)L	ASIV	ISIT	VIDI	A&DE	)DUK	VIEL
LEED	EKIS	ZIT)	TREL	M(LS	(T(P	*BWA	OSAM	*KIF	ITRA
R(GO	*LAD	UPAS	MEER	LALA	Z1S1	IM*L	NITT	I1SK	AFFO
OUTE	EEM*	D(LI	UTOM	LON&	WR:L	S(GZ	S(S*	DIR&	T*PI
DEMM	EELE	T(L*	M*BS	DEFW	))TE	EELA	*PR)	(BUK	IEV(
D)J+	RWLI	LOR(	VIT(	ES/F	DITA	(FIE	T(TA	I(SE	R*TR
OMUT	FOMI	KISU	AN(R	Z*S(	UIIE	I/KA	MES*	+DAN	PL)D
INER	M(PU	PRA*	T(NU	R*F*	EEPR	RILL	IS/A	ED)R	MWT)
NWRS	K*BL	(RI)	IL))	ISWA	Z)L&	EDEP	NERA	)R)F	WADE
/KSI	(SEK	OM(T	PL)&	ESED	)JET	*T))	ULID	AMAM	MIIV
UZ))	PR/T	R(RG	N)Z)	)+TR	PALO	J/S/	DESE	ID(M	)M(S
BTI1	ENU*	W7*T	DELA	DURO	WTAR	VWAF	EFIL	WLAT	))SI
Z+KI	AZED	ETT(	ZIZ1	/D*X	KIL)	S/LA	*SII	L)SE	OIS+
RAM(	Z1PA	(M(R	OEKA	JEKL	M*TI	KS(L	*LAN	S*SE	ENUD
AKIL	S*PL	R(BL	ORT(	T(RV	ROTT	(SEP	SASE	)TIZ	L(LI
LN)M	RV*J	LEEN	/TAX	RASE	EFAR	UNED	R(MM	RR(T	ABUF
KUNE	ONEE	IAJE	))VE	LAFI	N*VI	ARGA	ARS*	KAP+	KAFI
UIJE	ARTE	SS*T	VRID	SIR*	T/ST	RDED	D*GO	ITUR	(Z)D
ZUD(	DUN*	X(RP	K&(N	PAVW	WAPI	WSAR	T)PO	APII	V*RE
SERO	LM+S	LFR)	SEVI	DEDE	ELIR	VIRA	UDEV	RMAY	)M*S
F(Z(	NEDW	ITIK	(RT(	RT(F	(LET	E+KA	RTWT	TETR	ENN1
D(RE	WAJE	SEEV	D(LV	OLAK	EETW	APAR	FREE	UIRM	RENE
EB/L	SOR(	LUTR	AVOK	MIBO	(KT(	)T))	*XIZ	JEN(	S)ON
EPAJ	LELA	PAJ(	ARAT	GRAD	OR/S	A&KA	EDUP	MOM*	PALE
R(LE	Z/SU	KEPA	P(LI	SKOM	ALED	)M/1	FIST	(R)(	T)ZI
KARA	)TIM	WZUK	AVAS	)T)L	SIIR	RT*K	)LEP	WRTE	SIDE
IPRA	EMAG	KEEL	ELES	ABIT	PWRI	EM(R	+BLE	&DUI	IMAL
.D)P(	ASI)	IKR(	Z*KE	RTI)	NAPR	BLEP	A+SU	DOLE	/FL)

DAKJ	BU*)	A*Z(	M/E*	PEJV	PAI(	DEEE	M(IL	PRIN	ZEN*
TARE	Z1P*	R*I)	1DAR	SEV*	RUD*	S(N(	XARE	TZUR	MWAL
LIT*	X)R(	TAR(	MEG/	DEM(	O(LE	DARU	L*SE	ZIT)	SENE
KET)	PRAR	DERE	PITE	FEM)	GRWS	DRI1	APRA	N1P)	L*SE
DAPE	MAZ)	PROR	LID(	MUNE	K*TE	SITE	PAJE	KAT*	L(RE
S*TE	PASE	ZOT)	LIS(	P(KE	SWV)	PASE	L(KE	TRIL	R)RA
GALE	(KUI	D(TT	ZALE	VIKA	PWP*	NUS(	(LLW	(LLW	)RN(
S(S&	LOM*	X)LE	K(KK	P*TU	VWIY	KOI+	TEEE	L(Z/	SPIL
I*V*	MALE	NARE	IRI&	SEER	K(L&	FIV(	X*K)	DISI	K/I)
D)Z&	K/XE	GRAP	DISE	SAF+	PAT&	MOR(	R*TE	FOLL	G+IT
VJAR	SOM(	LETI	PWAR	REEE	DORA	SUI&	JAR)	PAV*	1D)&
UMME	UN(N	TEM*	M*TE	SUNE	UNAL	NAJ(	PRI&	TUI/	PER/
M(YA	UID)	VILE	NUIL	SWL&	ALLE	TIME	S(RE	)TI1	TET(
PLOR	O)R&	PROR	PAL*	ARS(	NAT(	FONE	PWRE	JARE	M(S(
TWAT	K/T1	F)T*	PLAL	PLUR	D)XE	TUNE	RINE	RFS)	P(NE
GRAR	SUDE	GAP*	DIT(	S(SI	JWR*	PAR&	MWAP	VIPL	T(V*
PLEE	PRIS	WBT*	KW7&	FIBI	B/TE	VWV*	K/D)	VENE	SAJ(
AVJR	ZIFF	*S)&	XER+	JWRE	PRAR	T*SI	R*XE	PWAR	VWAL
T)PL	BUF*	SI(Y	OM*J	PAJE	P(S&	D)V(	S(LE	SIBI	JU(V
TIJ)	DEOR	FI((	K/SI	M(I)	SURF	LUL(	BRI/	LIS)	PALE
TUDI	FENE	SELE	DIR(	RUR(	KWD(	SULE	PRI/	FIPL	LITI
MWRE	KOKA	N)N*	T+T&	TAU)	AVWA	TRAP	DEOR	VIT(	F)L+
TAKE	SID1	TR(R	DIL)	+SIF	+SI&	KIEE	BUPE	T*TI	DAPE
Z)RE	*VER	JAPA	R)SE	DAAL	Z)KA	POLL	BAP/	PONE	LAPE
SSOZ	JAYA	DAZ)	MASI	T)RE	L(P(	FUSA	S+SI	JUNE	(LG*
DOL)	T)P)	PR*S	PUI1	SIR(	KONE	BWT)	TRI+	+RT*	JUT&
MAV)	PAP*	JARE	AV(R	NAST	BRI/	AVWP	PR*X	*PT)	KLOR
UNNE	DET+	N)KU	JWTI	DEP(	ZUI1	XIAL	SISK	DISE	S+SA
KRAR	TUIE	K/PR	N(KE	TEOR	PAM)	XAT)	LUJ)	B+TE	ASIZ
N*DU	S*K&	ZOTE	SWV*	)TR)	KALE	L)KE	MAL&	*TRE	PAKE
SIME	GR(R	S(TR	VIEE	KW:(	BR(R	I EN(	PWAI	MORL	BUI+
SUS*	R(ME	RAR+	R*JE	N)P*	VORS	MOM*	UTTA	MORE	PLWA
NII/	MALE	N*TE	RENE	JABW	DIKE	REM*	M(S&	SEER	M(SE
R(R(	DOTI	MAK&	M(RA	TIY)	NWS&	)IRE	PASE	M*S&	BOLE
MWAR	PUKK	PAPE	FORE	EC'S)	DIF*	LIT(	K/S*	VEFE	LWV)
PLEE	X)PR	L*D*	PREE	NCKE	TWAR	SOT*	UNI/	LAJ*	MOPI
FRAR	PI(R	MONS	ADI)	ZWAL	XAL(	PAVA	PWAR	REI&	PRAR
BWAV	PR)&	DAS)	RULL	DITE	S*Y(	PARA	SUNE	OFOR	PRWA
M(R)	BAPR	V(T)	M(RE	KAS&	DIEF	R*SE	SEER	ZALE	MIRE
MAP&	O)ME	IERO	X(S*	MWAR	PAJ)	PR*S	PLIL	IDRE	L*SE
IEV)	LWJ&	(I*X	KL/J	PL)&	O)TI	REER	MUZE	PWV(	SUN)
B/T)	LENE	TURI	PRI&	JWAR	MWAR	Z)T(	L)FI	DAV(	ILE
K*T)	PR)&	PLER	VIKE	PASE	L)TE	VIRE	ZAF*	JABW	J*T)
SODE	N*F*	BROZ	SURE	+T)&	G+T&	BU)P	JAL(	S(ME	TR(D
L(R/	FI(N	KAR(	POLE	N(ST	SWAR	MODE	SIS&	PALU	K*RI
*TI/	RUS(	JURE	PEM*	K(LE	NUSI	ASAR	MWP)	LIT)	PUI)
PAP*	K(TI	SWAL	IREF	MERE	J+Z&	K(TE	O)T(	DELE	TZUD
NWBL	DAR*	KWRE	PWAR	Z)V(	PE(R	MUS&	BL(R	PWAP	PRI+
SUSE	M*V)	SOP(	R+Z7	TABE	SUI/	KISE	LUP*	DIVE	P/JE
R*XE	Z)SE	1T1)	PEN)	RET*	K/BW	MOT)	KR(L	D)*X	X)TE

Fig IV-26 Deux séries de 500 mots aléatoires, de longueur égale à 4 symboles phonétiques. Les mots de gauche sont composés à partir du comptage global reporté sur la figure IV-22. Les mots de droite sont établis à partir des mêmes statistiques, dans lesquelles les phonatomes initiaux, médians et finals ont été différenciés. Dans les deux cas les suites de trois consonnes ont été proscrites. Les mots aléatoires coïncidant avec des mots français sont soulignés.

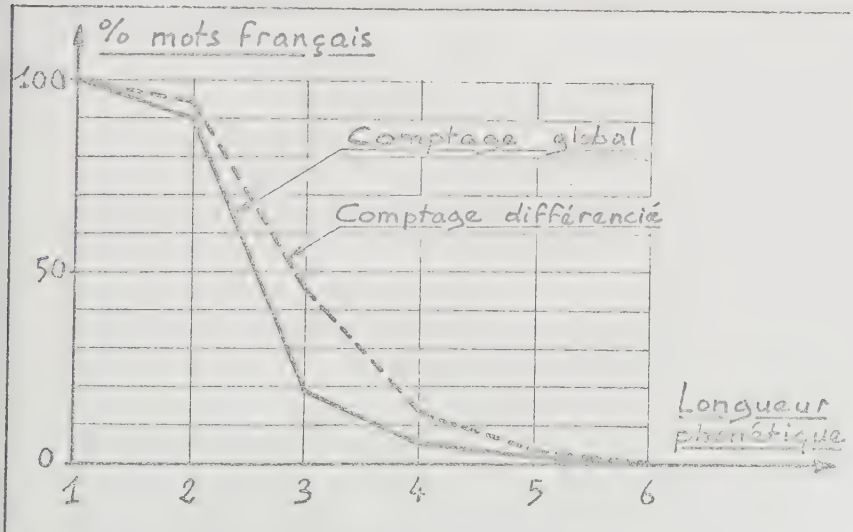


Fig IV-27 Nombre de mots français relevés dans des séries de 1000 mots aléatoires, en fonction de leur longueur phonétique. Le gain observé avec les comptages différenciés montre que des contraintes phonétiques supplémentaires ont été prises en compte.

Il est bien évident que la probabilité de répétition d'un même mot sur une liste de 1000 devient minime dès que n dépasse 3 .

De ces considérations se dégage une véritable méthode d'étude expérimentale des contraintes phonétiques : la recherche des mots français peut être automatisée si l'on dispose en mémoire, sous forme phonétique, de quelques milliers de mots français représentatifs qui constituent un corpus de référence. On peut simuler à volonté les contraintes phonétiques au niveau des probabilités de digrammes, ou au niveau de la sélection des mots aléatoires: par exemple tous les mots précédents ont été créés en interdisant les suites de trois consonnes. Un indice statistique objectif peut être déduit de la proportion des mots français créés.

IV.5 CONCLUSION DU CHAPITRE IV

Ce chapitre nous a permis d'examiner les structures de la parole sous divers aspects : phonologique, acoustique, phonétique. Nous savons maintenant que, si la signification est transportée par le squelette sémantique, celui-ci est éminemment variable d'un individu à l'autre, pour des raisons anatomo-physiologiques, et pour des raisons géographiques car la prononciation d'une même langue varie beaucoup selon les régions. Nous avons vu également que l'invariant du squelette sémantique est une forme temps-fréquence, et que pour compenser les différences entre locuteurs on peut effectuer diverses anamorphoses, dont la plus importante - et la plus simple - se réduit à un changement d'échelle fréquentielle. Le squelette sémantique peut sous certaines réserves se décomposer en une suite de formes acoustiques élémentaires appelées phonatomes, dont les symboles sont des digrammes phonétiques. Les phonèmes de la phonétique classique n'ont pas d'existence isolée dans le message parlé, qui est information, donc mouvement et transition; cependant la notation phonétique est parfaitement utilisable pour repérer les formes de la parole - phonatomes, mots ou phrases.

Les phonatomes existent effectivement de manière autonome : nous l'avons montré par l'analyse spectrographique et par l'étude de la répartition des digrammes, qui présente une stabilité remarquable et donne naissance à la notion de contrainte phonétique. La génération des mots aléatoires en est une conséquence directe, qui donne lieu à diverses applications. Avant de passer aux problèmes de la reconnaissance automatique, nous décrirons dans le chapitre suivant certains aspects de la synthèse de la parole qui démontrent l'efficacité des idées exposées jusqu'à présent.

V - LA SYNTHESE DE LA PAROLE

Dans ce chapitre, nous examinerons quelques aspects de la synthèse de la parole, dans la mesure où celle-ci permet de vérifier certaines des hypothèses précédentes sur les structures acoustiques de la parole. Nous devrons tout d'abord décrire les principaux moyens de synthèse, sans toutefois en faire une revue exhaustive, qui existe en bibliographie (bib. II-6, V-1, actes des congrès internationaux). Nous comprendrons au passage pourquoi des précurseurs comme KEMPELEN ou FABER ont pu rencontrer un succès mémorable. Nous présenterons ensuite la synthèse par phonatomes, telle qu'elle est réalisée au moyen des divers modèles d'Icophone. Nous envisagerons enfin les principales applications de la synthèse (réponse vocale automatique et transmission à faible bande passante), qui nous permettront au chapitre suivant de mieux situer le problème de la reconnaissance automatique.

V.1. LES OUTILS DE SYNTHESE

Nous distinguerons deux types de synthétiseurs, selon la démarche qui les a inspirés, démarche qui coïncide à peu près avec l'ordre chronologique. Les premiers visent à imiter le plus étroitement possible le mécanisme de la phonation; ce sont presque des modèles phonatoires. Les seconds visent à une synthèse efficace du message acoustique, sans référence à sa production.

V.1.1. Machines parlantes et modèles phonatoires

V.1.1.1. La machine de KEMPELEN

Jusqu'à la fin du dix-huitième siècle, les machines parlantes furent un prétexte à légendes et mystifications. Les premiers essais scientifiques - avec toute l'ambiguïté du terme à cette époque - peuvent être attribués à trois inventeurs dont l'un au moins - KEMPELEN - avait du génie. En France, l'abbé MICAL construisit une machine dont le fonctionnement était analogue à celui des boîtes à musique : ses deux "têtes parlantes" échangeaient en manière de dialogue les phrases suivantes :

- Le roi a donné la paix à l'Europe

- La paix fait le bonheur des peuples ....

Il semble que malgré ses imperfections, cette machine ait connu un certain succès. Un contemporain déclare que "... (les têtes parlantes) en disent assez pour qu'on ne puisse leur refuser le don de parole". (Bachaumont

...../

"Mémoires"). Il n'en reste guère qu'une description fort imprécise faite devant l'Académie des Sciences et signée par des personnages illustres (bib. V-2).

Au Danemark KRATZENSTEIN présenta une série de cinq résonateurs imitant les voyelles /a/, /œ/, /i/, /o/, /u/. Ces résonateurs de formes variées avaient à peu près le volume de la cavité buccale, et coloraient le spectre d'une anche libre de façon à produire les formants vocaliques.

La machine de KEMPELEN, gentilhomme de la cour d'Autriche-Hongrie, a le grand mérite d'exister encore - nous avons pu l'examiner au Deutsches Museum de Munich - et d'avoir fait l'objet d'une description suffisamment précise pour en permettre la reconstitution.

Dans son livre (bib. V-3), KEMPELEN explique l'échec de sa première machine, dans laquelle divers résonateurs étaient mis en action au moyen des touches d'un clavier. La machine reproduisait correctement certaines voyelles et certaines consonnes, mais il était impossible de relier ces divers sons entre eux. KEMPELEN se remit à l'ouvrage, en posant comme principe qu'une machine parlante ne devait avoir qu'"une seule glotte et une seule bouche".

La machine définitive, que nous avons reconstituée à partir de ses descriptions (bib. V-4), présente une grande simplicité fonctionnelle (fig. V-1, V-2, V-3). L'opérateur appuie son coude droit sur le soufflet; sa main droite commande les divers leviers et ouvertures qui se trouvent sur le dessus de la boîte à vent. A l'intérieur de celle-ci se trouvent une anche qui vibre quand la pression d'air est suffisante, et des soupapes commandées par des leviers, qui permettent à l'air de passer à travers les sifflets du /s/ et du /ʃ/. Le /r/ est produit par le 3e levier, qui amène une aiguille de laiton au contact de l'anche, introduisant ainsi des perturbations dans son fonctionnement; il en résulte une sorte de râclement qui imite le son /r/.

Les voyelles sont obtenues en faisant varier avec la main gauche les caractéristiques du résonateur en caoutchouc qui constitue la "bouche". Complètement ouverte elle donne un son voisin de /a/ ou /œ̃/. A moitié fermée on peut entendre /œ/ ou /y/; presque complètement fermée on entend /u/. Tous ces sons vocaliques proviennent du déplacement d'une seule zone formantique, peu accentuée, entre 1000 et 2000 Hz.

La bouche a également pour fonction la production de certaines consonnes comme le /p/. A cet effet on commence par obturer complètement. En continuant à appuyer sur le soufflet on provoque une surpression dans la boîte à vent et dans la bouche, qui communiquent par le moyen d'un tuyau de faible diamètre (by-pass). En écartant brusquement la main gauche on produit alors le son recherché, qui s'enchaîne avec la voyelle suivante. Le petit soufflet additionnel renforce l'efficacité de la surpression. Les sons /b/, /t/, /d/, /k/ n'existent pas dans la machine : KEMPELEN les remplaçait sciemment par un /p/ plus ou moins rapide.

On peut aussi, en jouant sur les sifflets, la bouche et les narines, fabriquer d'autres sons, comme /m/, /l/, /f/, /v/, etc... Tous ces sons, considérés isolément, sont plus ou moins bien imités. Mais l'essentiel est qu'ils puissent s'enchaîner les uns aux autres et composer ainsi des mots compréhensibles. On arrive assez facilement à fabriquer les mots "papa", "maman" "alors", "ça va", etc..., compréhensibles d'emblée par des auditeurs non prévenus. A vrai dire, cette machine est un véritable "instrument de parole" dont il faut apprendre à jouer. KEMPELEN en était certainement un virtuose,

..../

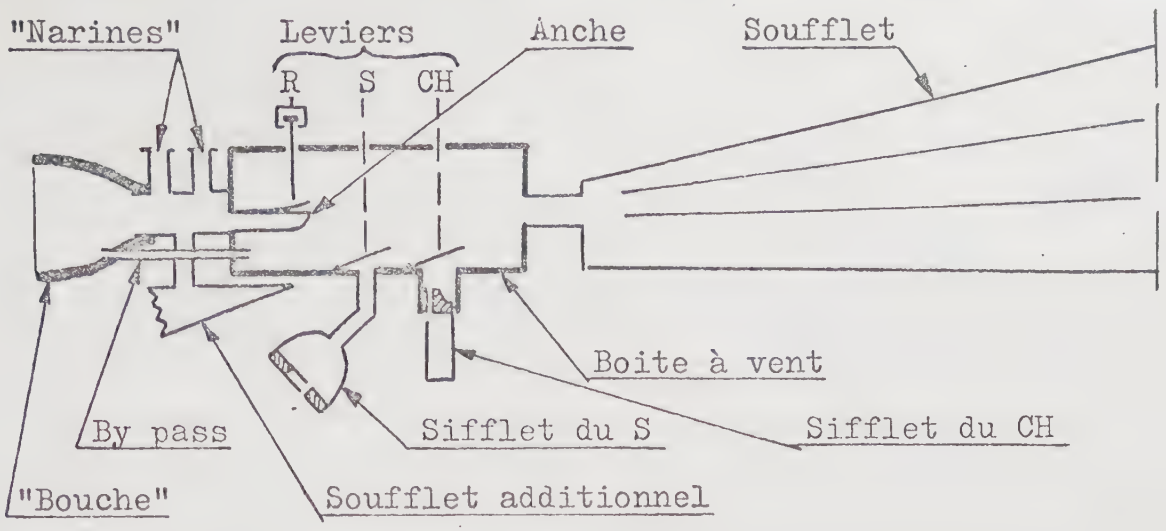
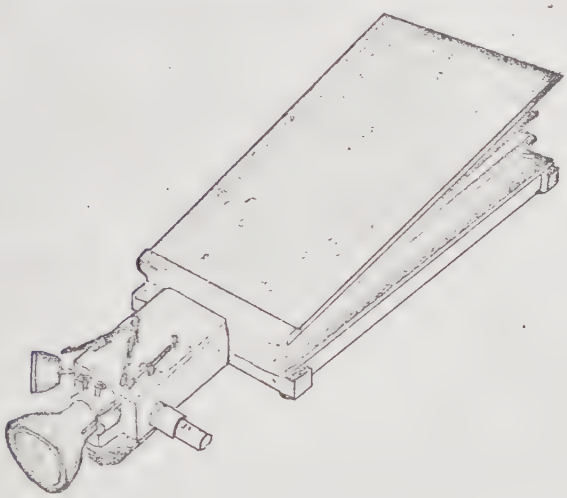


Fig V-1

Schéma fonctionnel de la machine de KEMPELEN.

Fig V-2

Reconstitution de la machine de KEMPELEN, d'après les données de son livre.



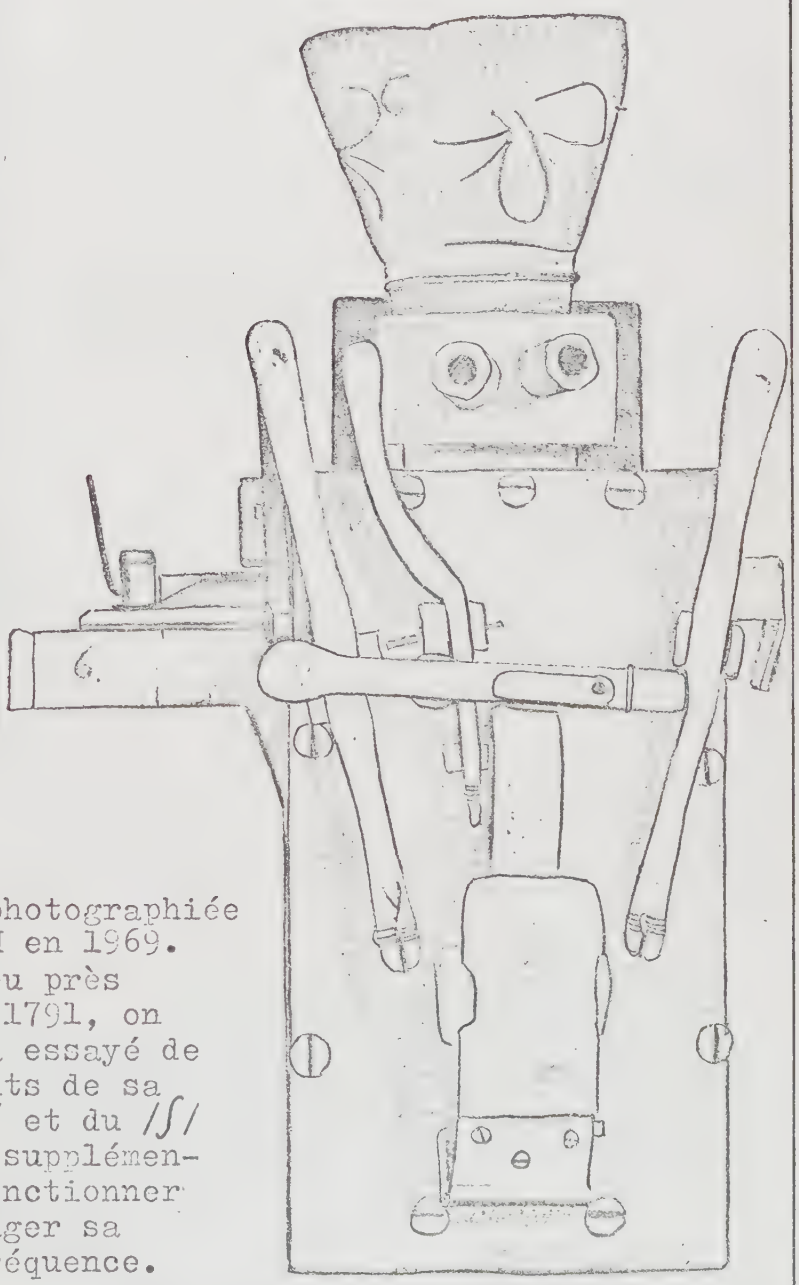


Fig V-3

La machine de KEMPELEN, photographiée au Deutsches Museum de MUNICH en 1969. Bien que l'ensemble soit à peu près conforme à la description de 1791, on peut constater que l'auteur a essayé de remédier aux principaux défauts de sa machine : les sifflets du /s/ et du /ʃ/ sont modifiés ; deux leviers supplémentaires permettent de faire fonctionner l'anche à volonté, et de changer sa longueur vibrante, donc sa fréquence.

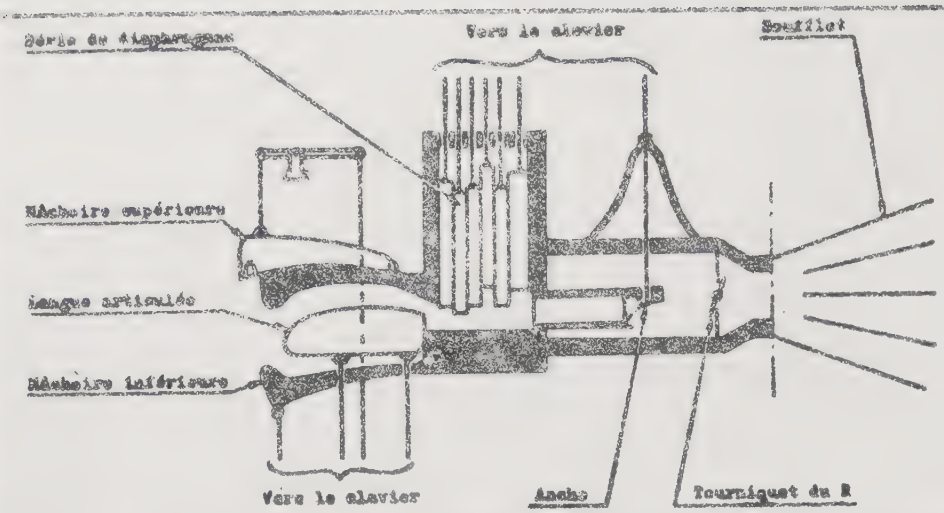


Fig V-4

Schéma fonctionnel de la machine de FABER (1835).

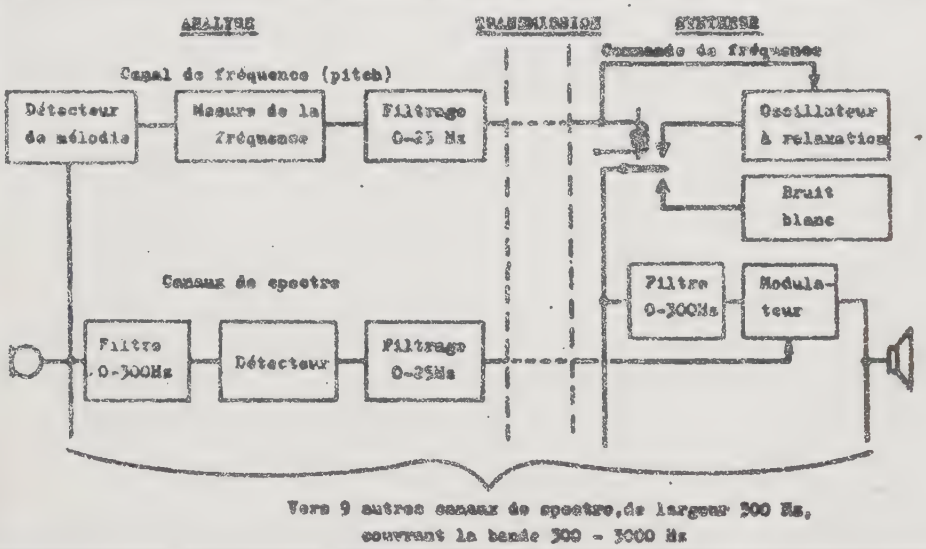


Fig V-5

Schéma du VOCODER à canaux (DUDLEY 1939).

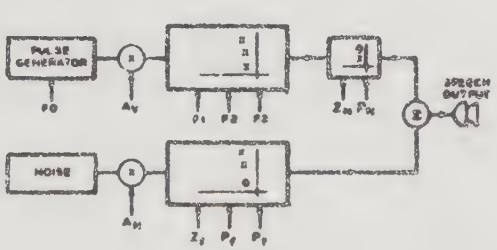


Fig V-6

Principe d'un synthétiseur paramétrique digital (FLANAGAN).

-Fo et Av : fréquence et amplitude du spectre de raies (voix)

-F_1, F_2, F_3 fréquences des trois premiers formants.

-Zn, Pn : caractéristiques spectrales des semi-voyelles.

-An : amplitude du spectre continu.

-Zf, Pf : caractéristiques spectrales des consonnes.

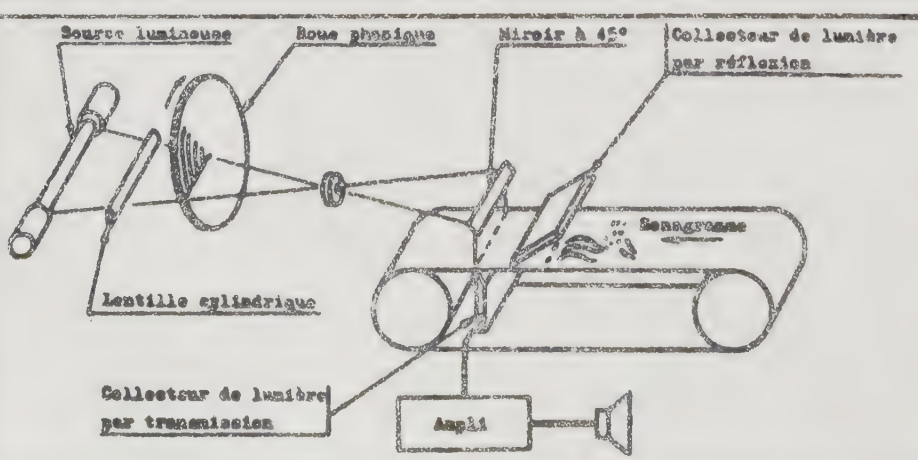


Fig V-7

Principe du Pattern Playback de COOPER (1951).

et nous pouvons le croire lorsqu'il écrit :

"... Toute imparfaite qu'elle soit, elle donne du moins de bons principes pour en construire une plus parfaite ... je lui fais prononcer .... plusieurs centaines de mots clairement et distinctement. Par exemple : Papa, Maman, Marianna, Roma, Maladie, Santé, Astronomie, Anatomie, Chapeau, Racine, Supé, Charmante, Opéra, Comédie, Pantomime, et aussi des mots longs et difficiles, tels que Constantinopolis, Monomotapa, Mississipi, Astrakan, Anastasius, etc... "

L'étude des témoignages laissés par les contemporains montre que KEMPELEN utilisait aussi la suggestion : en général les mots prononcés par la machine avaient été demandés par l'assistance; ils étaient donc déjà dans l'esprit des auditeurs, qui les comprenaient d'autant mieux. KEMPELEN avait conscience de ce phénomène absolument fondamental dans le processus de la compréhension :

"... On est surtout aisément induit en erreur quand on sait d'avance le mot que la machine doit dire, et lorsqu'elle le prononce on s'imagine l'avoir entendu ..."

KEMPELEN avait compris l'importance de la liaison des sons entre eux, et de la suggestion ou prévisibilité, qui constituent deux points fondamentaux de la communication parlée. Mais dans sa simulation des organes phonatoires la langue, les lèvres et les dents étaient oubliés.

V.1.1.2. La machine de FABER

Joseph FABER présenta en 1935 à Vienne une machine plus élaborée, qui parlait à voix haute et à voix chuchotée, et qui chantait "God Save the Queen" ! Elle possédait un grand nombre de perfectionnements par rapport à celle de KEMPELEN (fig. V.4).

- les deux maxillaires et la langue sont articulés,
- le profil de la "glotte" est variable selon la position d'une série de diaphragmes coulissants,
- une rasette mobile permet de faire varier la longueur de l'anche, et donc l'intonation de la voix,
- la commande de la machine se fait au moyen d'un clavier comportant 14 touches.

Cette machine a sans aucun doute donné des résultats intéressants. Il ne nous en reste malheureusement aucune trace : après avoir été montrée pendant un demi-siècle dans les capitales du vieux et du nouveau monde, elle a, selon toute vraisemblance, terminé sa vie dans les caves de l'Ecole de Médecine de Paris.

Après FABER, il semble que la synthèse de la parole ait régressé. De nombreux physiciens se sont attaqués au problème, mais ils l'abordèrent de manière analytique en s'intéressant surtout aux voyelles. Ils abandonnèrent ainsi la méthode empirique et globale qui avait fait la fortune de KEMPELEN et de FABER. Pour imiter les voyelles HELMHOLTZ utilisa des combinaisons de diapasons, KOENIG des sirènes, MILLER et STUMPF des tuyaux d'orgue, MARAGE et PAGET des résonateurs, etc... Il fallut attendre 1937 pour que RIESZ présentât une machine parlante qui reprenait dans ses grandes lignes celle de

...../

FABER. Maniée par un opérateur habile, cette machine produisait, paraît-il, une parole courante et compréhensible. Ce fut, à notre connaissance, le dernier en date des synthétiseurs mécaniques.

V.1.1.3. Le VODER

Le VODER (VOice DemonstratoR) a été présenté en 1939 par DUDLEY, un ingénieur de la compagnie Bell Telephone. C'est un banc de 10 filtres passe-bande répartis entre 300 et 3000 Hz et connectés en parallèle. Les circuits peuvent être excités soit avec du bruit blanc, soit avec un signal très riche en harmoniques dont la fréquence est commandée par une pédale. Chaque touche commande l'amplitude du signal dans l'une des bandes; trois touches supplémentaires permettent de reproduire les plosives.

Comme pour toutes les machines précédentes, il est difficile d'estimer objectivement la qualité de la parole produite. La fréquence des formants n'est définie que par échelons de 300 Hz, et le choix entre bruit et spectre harmonique se fait en tout ou rien : la nécessaire continuité de la parole n'est maintenue que par le "jeu" de l'opérateur, qui doit être extraordinairement habile.

V.1.1.4. Les analogues électriques du conduit vocal

Dans ces appareils, le conduit vocal est considéré comme une suite de tubes cylindriques élémentaires, de section et longueur variables. Chaque élément présentant une certaine impédance acoustique complexe, on lui fait correspondre un circuit électrique d'impédance analogue. En ajustant les paramètres de chaque circuit et en excitant le montage en divers points par un bruit ou un son harmonique, on simule tous les sons permanents ou semi-permanents de la parole. Ce processus peut évidemment être contrôlé par calculateur. Un système de ce type mis au point par C.H. COKER aux laboratoires Bell fournit même sur oscilloscope un affichage du profil du conduit vocal, avec ses modifications pendant les transitions de la parole.

Bien que ces appareillages soient très intéressants sur le plan de la synthèse, nous reprendrons à leur sujet les critiques déjà formulées à l'encontre des théories articulatoires : les paramètres articulatoires sont des grandeurs très mal définies et les phénomènes importants acoustiquement correspondent souvent à des détails à peine visibles. Enfin, la simulation électrique d'un phénomène acoustique est par trop simpliste : comment rendre compte électriquement des modifications spectrales dues aux écoulements aériens, à l'amortissement des tissus, etc.. tous phénomènes qui se réalisent en trois dimensions, avec des vibrations d'une autre nature ? En ce qui concerne la simulation de l'appareil phonatoire, il nous semble que KEMPELEN était plus près de la réalité.

V.1.2. Les synthétiseurs de parole

V.1.2.1. Le VOCODER

En même temps que le VODER, DUDLEY présentait le VOCODER (VOice CODER), qui réalisait l'analyse et la synthèse instantanées de la parole. Le but de

...

l'opération était de réduire la redondance, donc la bande passante de la parole, afin d'accroître le rendement des transmissions téléphoniques.

L'organe d'analyse du VOCODER à canaux est un banc de 10 filtres passe-bande, identique à celui du VODER (fig. V-5). Le message à transmettre est ainsi décomposé en dix bandes de largeur 300 Hz, qui après détection et filtrage passe-bas donnent lieu à dix signaux utilisés pour commander un synthétiseur comparable à celui du VODER. Un onzième canal achemine un signal indiquant, lorsque le message est voisé, la fréquence du spectre de raies ("pitch"). Le bilan informationnel est séduisant : au lieu des 3100 Hz requis par le téphone, on utilise seulement une bande de 11 x 25 = 275 Hz (soit environ 2000 bits par seconde), ce qui permet en principe de multiplier par dix le nombre de communications possibles sur une seule ligne.

Cependant l'intelligibilité de la parole synthétique est demeurée médiocre (du moins avec ce taux de compression). Ce demi-échec a une double origine :

- Le squelette sémantique du message est quantifié trop grossièrement : la fréquence des formants n'est définie que par échelons de 300 Hz, et la limitation à 25 Hz par canal supprime les phénomènes brefs de la parole dont nous avons plus haut montré l'importance.

- La distinction voisé-non voisé est en réalité beaucoup plus floue qu'il ne paraît à première vue, sur la foi des analogies électriques de la phonation : elle dépend du locuteur et du contexte phonétique. Souvent les deux types de spectre coexistent dans les mêmes sons; la raison en est simple : si le spectre de raies provient nécessairement de la glotte, le spectre continu peut avoir son origine à tous les niveaux du conduit vocal, glotte comprise. Un système décidant si le son est voisé ou non est par définition sujet à beaucoup d'erreurs et, fait plus grave, ce taux d'erreurs est variable selon les voix. Par ailleurs, la détection du "pitch" est rendue difficile par la diversité des timbres individuels : l'énergie des premiers harmoniques est très variable et les détecteurs font très souvent des erreurs d'octave.

Ces considérations permettent de comprendre l'immense effort de recherche entrepris depuis 1939, surtout aux Etats-Unis, pour améliorer le VOCODER : on a construit d'innombrables variantes du VOCODER de DUDLEY, et d'innombrables modèles de détecteurs de pitch. Il semble que seul le VOCODER "à bande de base" (voice-excited vocoder) donne de bons résultats : dans cet appareil, on transmet intégralement les fréquences comprises entre 300 et 1000 Hz environ qui contiennent l'essentiel de l'information relative au pitch et au voisement. Naturellement, le taux de compression est moins élevé qu'on ne pouvait l'espérer (de l'ordre de 3, au lieu de 10), mais la qualité obtenue est sensiblement celle du téléphone ordinaire.

Il semble bien que, après avoir été 30 ans durant le moteur des recherches sur la parole, le VOCODER soit en voie d'abandon, le bénéfice attendu étant maintenant jugé peu rentable sauf dans quelques cas particuliers. La technologie actuelle des télécommunications (transistor, satellite, laser, etc...) donne en effet l'accès à des bandes passantes beaucoup plus étendues qu'autrefois.

...../

V.1.2.2. Les synthétiseurs paramétriques

La synthèse paramétrique (ou "à formants") cherche à reconstituer la parole à partir de ses paramètres, définis par une analyse préalable. Les paramètres les plus utilisés sont la fréquence Fo du fondamental, les fréquences F_1 F_2 F_3 des trois premiers formants, l'amplitude du spectre de raies, l'amplitude du bruit, etc... (fig. V-6). Chacun des paramètres est exprimé en fonction du temps, par une courbe ou par une suite de nombres. On peut ainsi faire varier chaque paramètre isolément, et étudier le résultat auditif (bib. V-15).

La qualité de la voix produite est excellente. C'est l'outil de choix des phonéticiens. L'inconvénient est la nécessité d'une analyse préalable : pour synthétiser une phrase, il faut d'abord en faire le sonagramme, relever l'évolution de tous les paramètres avec précision et les convertir en données assimilables par le synthétiseur, ce qui demande beaucoup de temps.

Récemment apparue, la synthèse par règles (synthesis by rules) tente de faire élaborer par un calculateur les transitions des paramètres entre phonèmes successifs : la donnée d'une suite de symboles phonétiques entraîne le calcul de l'évolution temporelle des paramètres et leur transfert vers le synthétiseur. Après nombre d'essais, il semble que cette synthèse soit maintenant au point aux laboratoires Bell (bib. V-5). Des études sont également menées en France sur cette question (bib. V-16).

V.1.2.3. Les relecteurs de sonagrammes

Le principe de ces appareils consiste à "relire" le sonagramme, c'est-à-dire à transformer en signal sonore le sonagramme schématisé sur un support transparent.

a) Le play-back de COOPER (1950)

Cet appareil (fig. V-7) se compose d'une roue en verre sur laquelle sont disposées 50 pistes concentriques semblables aux pistes sonores des films cinématographiques. Le faisceau lumineux émis par la source se trouve divisé en 50 pinceaux dont le premier est modulé à 120 Hz, le second à 240 Hz, etc... jusqu'à 6000 Hz. Ces pinceaux sont eux-mêmes modulés par le sonagramme, et le résultat est recueilli soit par transparence, soit par réflexion, au moyen d'une cellule phoélectrique. La voix produite est "recto-tono" sur un fondamental de 120 Hz, et possède une intelligibilité surprenante, compte tenu de la simplicité des moyens mis en oeuvre.

Le play-back, utilisé par des psychologues et des linguistes des laboratoires HASKINS (dont LIBERMANN et DELATTRE), est à l'origine de la théorie du "Locus" qui a connu ces dernières années un grand succès en phonétique, et de la théorie motrice de la perception, selon laquelle le message parlé est perçu par reconnaissance des mouvements articulatoires. Nous avons déjà exposé notre position sur ces deux points, aussi n'y reviendrons-nous pas. Nous remarquerons seulement deux caractéristiques essentielles du play-back :

- L'appareil ne semble pas apte à reproduire les phénomènes de très courte durée (clics ou attaques de quelques millisecondes) qui transportent pourtant une part notable de l'information sémantique.

...../

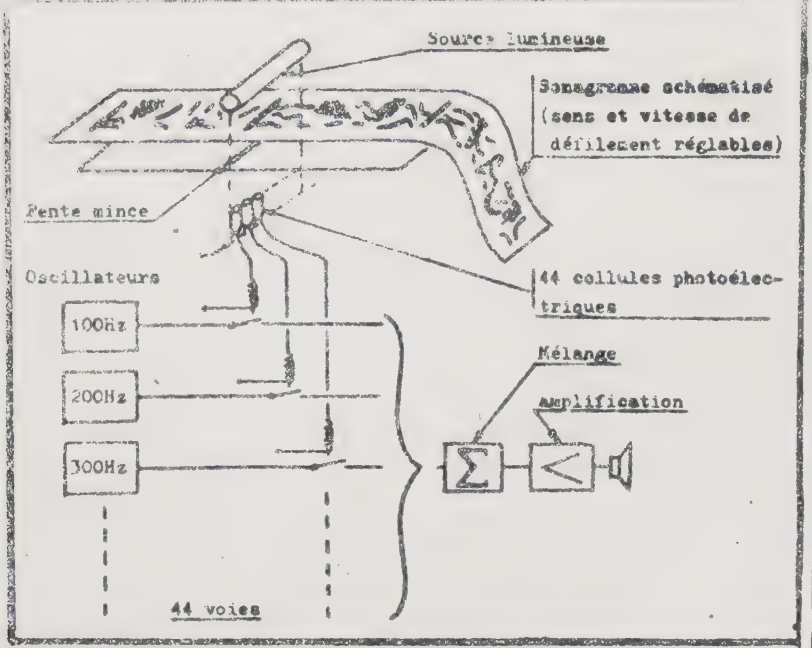


Fig V-8 (ci-dessus) - Principe de l'Icophone II, à commande optique

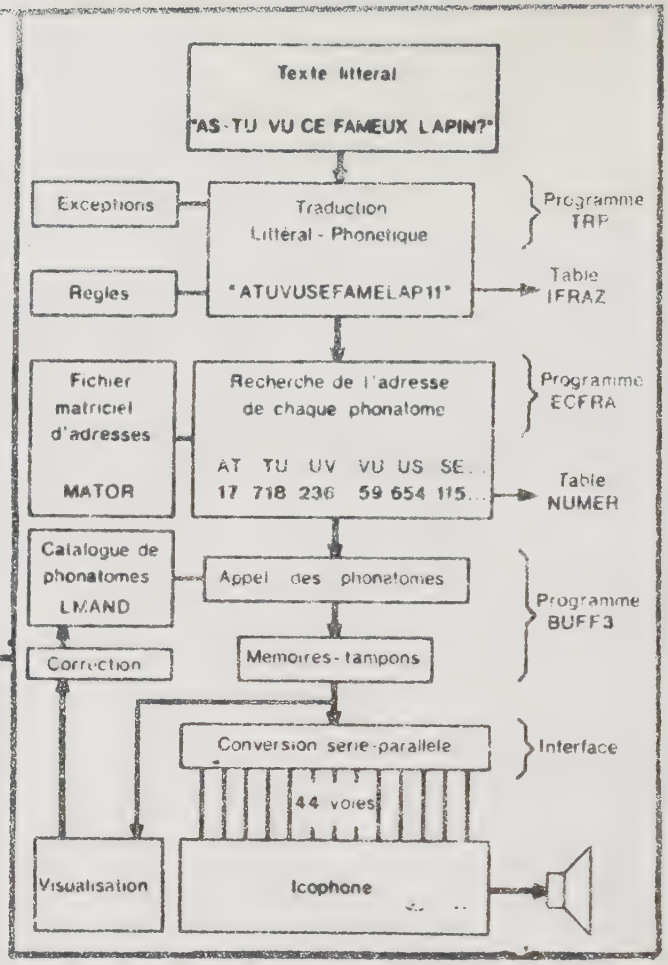
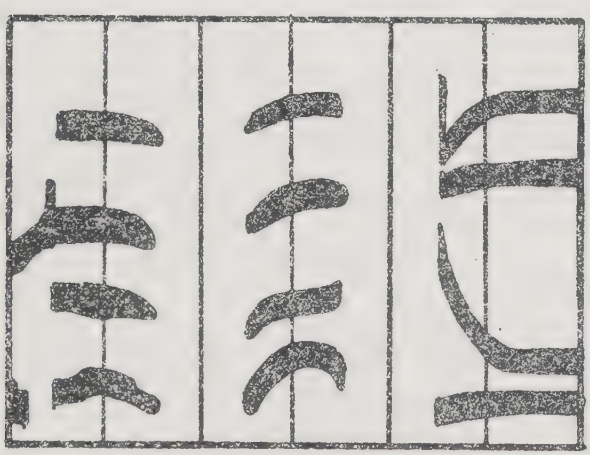


Fig V-9 (ci-contre) - Principe de l'Icophone III



LE EB BA AT T+ ++

LE EB BA AT T+ ++

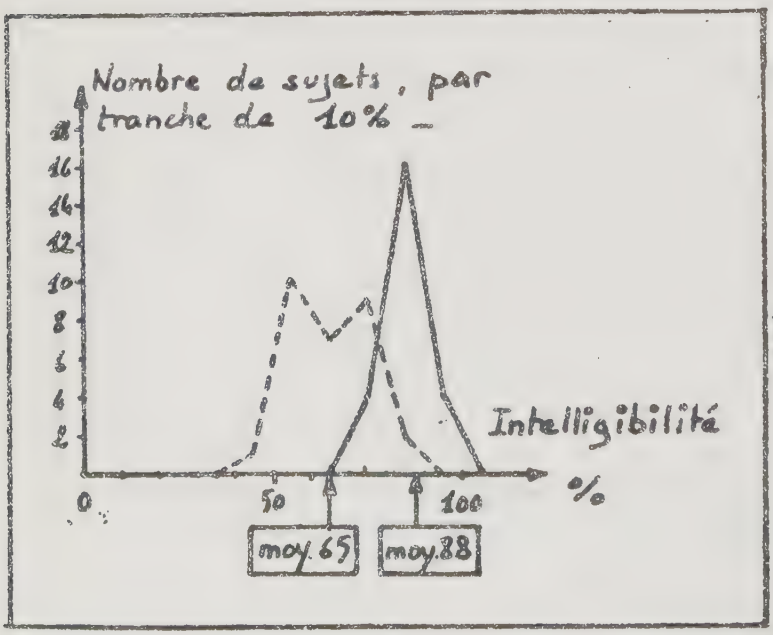


Fig V-10 (ci-dessus) - La même expression ("le bateau"), avant et après le voisement du dictionnaire

Fig V-11 (ci-contre) - Résultats des tests d'intelligibilité :
- en trait interrompu : 11-3-7
- en trait plein : 24-11-70

- L'intensité sonore de chaque fréquence est proportionnelle à la largeur du trait.

Nous allons voir que sur ces deux points l'Icophone présente des différences importantes.

b) L'Icophone II à commande optique

L'Icophone est un type de synthétiseur conçu et réalisé au département de Mécanique de l'Université PARIS VI (Laboratoires d'Acoustique et d'Electronique). Le modèle I, construit en 1965, utilisait l'infrastructure du Sonagraphe. Le principe du modèle II est extrêmement simple (fig. V-8) : la fréquence sinusoïdale issue d'un générateur est envoyée dans le circuit de mélange lorsqu'une zone opaque du sonagramme schématisé défile devant la cellule correspondante. Les cellules fonctionnent par tout ou rien. L'indépendance des générateurs permet d'éviter, si on le désire, l'harmonicité des composantes. Le dessin en hachures permet d'introduire du bruit au lieu de fréquences pures : on fabrique ainsi à volonté de la parole voisée recto tono ou de la parole chuchotée. La réponse temporelle de l'appareil a été particulièrement soignée : des phénomènes de durée inférieure à 5 ms peuvent être reproduits en synchronisme sur les 44 voies. C'est sur cet appareil qu'ont été effectuées les recherches du laboratoire relatives à la théorie gestaltiste de la parole (bib. IV-14, IV-12, V-7).

V.2. LA SYNTHESE PAR PHONATOMES

Nous avons au chapitre IV mis en évidence l'existence acoustique et phonétique des phonatomes. Nous venons de voir que, pour être intelligible, la parole synthétique devait respecter en premier lieu l'intégrité du squelette sémantique considéré comme une Forme. Une solution efficace au problème de la synthèse consiste à connaître les squelettes de tous les phonatomes possibles, et à les assembler comme des dominos. Ainsi le mot "Icophone", /ikɔfɔn/ peut être reconstitué en juxtaposant les phonatomes /ik/, /kɔ/, /ɔf/, /fɔ/ et /ɔn/. Des méthodes analogues, basées sur la notion de phonatome (ou "diphone" dans la terminologie américaine) ont été présentées en 1968 par DIXON et MAXEY (bib. V-8), SAITO et HASHIMOTO (bib. V-9), et utilisaient toutes deux un synthétiseur paramétrique commandé par un calculateur.

La mise en oeuvre de cette méthode sur l'Icophone s'est faite en plusieurs étapes que nous allons détailler quelque peu.

V.2.1. Etablissement d'un premier dictionnaire sur l'Icophone à commande optique.

L'Icophone II permet de transformer en parole un sonagramme schématisé, réduit à son squelette sémantique. La difficulté réside justement dans cette schématisation, d'autant plus ardue que la séquence à synthétiser est de plus courte durée (absence de signification).

...../

V.2.1.1. Nombre de phonatomes utiles

Ayant choisi un ensemble de 30 phonèmes, nous avions à schématiser 30 x 30 = 900 phonatomes. En réalité, l'étude statistique évoquée plus haut a montré que 600 seulement sont nécessaires. De plus, tous les phonatomes ne comportant pas de plosives peuvent être en première approximation considérés comme réversibles temporellement : c'est-à-dire que /ar/, par exemple, a la même représentation que /ra/, au sens de défilement près. Le nombre de phonatomes à schématiser se réduit ainsi à 400 environ.

V.2.1.2. Schématisation des phonatomes

Les premiers schémas de phonatomes ont été élaborés par un processus semi-empirique d'analyse et synthèse demandant un certain entraînement : les phonatomes sont enregistrés par l'opérateur en voix chuchotée ou voisée (souvent de plusieurs manières), analysés au moyen du Sonagraphe de façon à faire émerger au mieux les formes utiles, schématisés manuellement sur la bande transparente de l'Icophone, écoutés et retouchés jusqu'à ce que l'intelligibilité soit satisfaisante. A ce point, l'opérateur se heurte au phénomène de suggestion que KEMPELEN mettait à profit, et qui, ici, se révèle extrêmement gênant il est indispensable de se soumettre à l'avis d'auditeurs non prévenus, même lo que le schéma semble avoir atteint une perfection évidente pour l'opérateur (bib. IV-16).

La schématisation s'accompagne d'une normalisation dimensionnelle : chaque phonatome est inscrit dans un rectangle de 20 x 88 mm correspondant à une durée d'environ 80 ms (variable selon la vitesse de défilement) et à l'intervalle 100 - 4400 Hz.

V.2.1.3. Comparaison des dictionnaires

Nous avons vu plus haut que les différences entre les voix pouvaient être considérables. Sachant cela, les trois chercheurs du Laboratoire d'Acoustique avaient décidé de fabriquer chacun son dictionnaire personnel et de l'amener à un taux d'intelligibilité suffisant pour que leur comparaison pût être fructueuse. L'ensemble de ce travail n'est pas terminé, le dictionnaire le mieux au point actuellement étant celui de Melle CASTELLENGO.

Il est néanmoins très édifiant de comparer quelques phonatomes des trois dictionnaires, phonatomes qu'un auditeur qualifié et impartial (Mme BOREL-MAISONNY) a jugés également intelligibles (fig. V-12). Ici encore nous pouvons estimer la difficulté de la reconnaissance automatique appliquée à un locuteur quelconque : d'un locuteur à l'autre apparaissent des différences considérables, non seulement sur la position des formants (le dictionnaire de Melle CASTELLENGO possède d'ailleurs par construction une anamorphose fréquentielle régulière de l'ordre de 0,8), mais aussi sur la forme acoustique objective de chaque phonatome. Tout se passe comme si chaque locuteur s'efforçait, avec ses moyens phonatoires propres, de fabriquer des phonatomes ressemblant suffisamment aux phonatomes-types de la langue utilisée, cette approximation étant jugée suffisante quand elle permet d'éviter toute confusion entre phonatomes voisins. Nous avons ici une illustration de la perception catégorielle évoquée au chapitre III.

...../

Fig V-12

Schématisation des phonatomes de /ara/ et /ãnã/ par trois expérimentateurs différents. L'intelligibilité de ces schémas, lus par l'Icophone, a été jugée également satisfaisante par un auditeur impartial. Cependant les différences sont nombreuses d'un expérimentateur à l'autre, et montrent que les formes optima ne sont pas encore atteintes. On notera que les schémas de gauche correspondent à une voix chuchotée, et qu'il existe entre ces trois voix synthétiques des anamorphoses fréquentielles et spectrales.

V.2.2. Utilisation de l'Icophone III à commande numérique

L'Icophone optique est d'un usage relativement souple en ce qui concerne la schématisation des phonatomes isolés. Par contre, la composition de messages entiers est extrêmement fastidieuse : il faut environ 10 minutes pour assembler et recopier les phonatomes d'une courte phrase. La commande numérique de l'Icophone permet d'améliorer considérablement le processus de synthèse, en qualité et en rapidité.

V.2.2.1. L'Icophone III à commande numérique (brevet ANVAR n° PV 182 925 du 31 décembre 1968, enregistré définitivement le 22 février 1971 sous le n° 1 602 936).

Le développement de cet appareillage a été réalisé au Centre de Calcul Analogique du C.N.R.S. (bib. V-10, V-11, IV-12, V-17). L'ordinateur utilisé est un I.B.M. 1130 muni de ses périphériques classiques et d'une sortie canal qui permet d'utiliser des périphériques non normalisés. A la mémoire centrale de 32 000 mots de 16 bits est associée une mémoire sur disque de 512 000 mots.

L'appareil transforme en parole intelligible l'information contenue dans un texte phonétique ou orthographié (fig. V-9). Ce dernier est d'abord traduit en une suite de symboles phonétiques, qui permet d'appeler les phonatomes correspondants situés sur le disque; puis ceux-ci sont juxtaposés et édités sur l'Icophone III.

a) Traduction phonétique (programme T R P de D. TEIL)

Le message parlé étant continu, il est nécessaire de transformer le texte littéral, constitué de mots séparés et orthographiés, en une suite ininterrompue de symboles phonétiques, dont certains sont redoublés afin d'introduire un rythme vraisemblable.

La prononciation est régie par des règles et des exceptions : chaque mot est comparé à l'ensemble des mots présentant une prononciation exceptionnelle, par exemple :

" oignon" /ɔ ɲɔ̃/, "psychologie" (psikɔlɔʒi/ .

Si le mot à traduire ne se trouve pas dans cette liste, on admet que sa prononciation se fait selon les règles, qui sont nombreuses mais relativement faciles à programmer. Le mot est examiné lettre par lettre, et soumis à des règles telles que :

- "o" suivi de "i" se prononce /ua/, (ou "WA" en code machine),
- "c" suivi de "h" se prononce /ʃ/ ("X" en code machine) etc...

En ce qui concerne le rythme, une règle rudimentaire et néanmoins satisfaisante consiste à doubler ou tripler la durée de la voyelle supportant l'accent tonique, qui est en général la dernière voyelle (sauf /ə/ muet) avant une ponctuation.

b) Recherche des phonatomes en mémoire

Dans une phase préliminaire, les phonatomes du dictionnaire ont été quantifiés en 20 colonnes et 44 lignes, et transférés sur cartes perforées.

..../

Ce fichier est stocké en mémoire de masse (disque), l'adresse étant repérée au moyen d'une table matricielle (MATOR). Le rangement sur disque est tel que les phonatomes les plus fréquents soient accessibles dans le temps le plus court (utilisation des statistiques établies au chapitre IV).

Pour prononcer le mot "Paris", qui se présente sous la forme "PARII" après traduction phonétique, on cherche dans MATOR l'adresse du phonatome "PA", qui se trouve à l'intersection de la ligne P et de la colonne A, puis les adresses des phonatomes "AR", "RI", et "II". Le message est alors constitué d'une suite d'adresses (table NUMER). Les phonatomes sont ensuite extraits de la mémoire de masse, sous forme de séquences binaires.

c) Régulation temporelle

Pour assurer la continuité de la séquence vocale, les phonatomes doivent être transmis à l'Icophone sans interruption, avec un débit régulier et réglable. Il est nécessaire d'intercaler entre le disque et l'organe de gestion du synthétiseur (sous-canal TITN) un système de trois mémoires-tampons, de telle sorte que l'une des trois soit toujours pleine, et prête à se décharger sur l'Icophone.

d) Sortie vocale

Dans l'Icophone III les cellules photoélectriques sont remplacées par des portes électroniques dont la fermeture est commandée par des signaux binaires. Il est toutefois nécessaire d'effectuer un décodage de ces signaux, car l'information destinée aux 44 oscillateurs se présente sous la forme d'une seule séquence d'impulsions, alors que la mise en circuit des oscillateurs doit pouvoir être simultanée, et maintenue pendant toute la durée séparant deux "événements".

e) Visualisation

Indépendamment de la sortie vocale, les phonatomes constituant le message peuvent être affichés sur un tube à mémoire géré par un autre sous-canal du même ordinateur. Diverses opérations sont possibles, telles que l'agrandissement d'une partie de l'image, la modification ponctuelle ou globale d'un phonatome, son rangement en mémoire, l'écoute immédiate du phonatome modifié et non modifié, l'écoute de phonatomes isolés ou associés à d'autres pour former des mots, etc...

L'Icophone III constitue donc un puissant moyen d'étude expérimentale des structures acoustiques de la parole. Melle CASTELLENGO l'a amplement utilisé pour améliorer le premier dictionnaire, dans le cadre du contrat DRME n° 69/195.

V.2.2.2. L'amélioration du dictionnaire

a) Le voisement

Le premier dictionnaire introduit en machine avait été établi au moyen de la voix chuchotée, qui fournit des formes temps-fréquence très nettes. Cependant la mauvaise discrimination entre consonnes sourdes et sonores (/p/ et /b/, /t/ et /d/, /s/ et /z/ ...) entraîne certaines ambiguïtés qui

...../

peuvent être gênantes si les mots utilisés sont inconnus de l'auditeur.

Le principe de l'Icophone permet de passer très simplement de la voix chuchotée à la voix "voisée", en excitant les oscillateurs dont les fréquences se trouvent au-dessous du premier formant lorsqu'il existe, et en laissant fonctionner les 2 ou 3 premiers oscillateurs pendant toute la durée des consonnes sonores (fig. V-10).

Le voisement implique l'accord rigoureux des dix ou quinze premiers oscillateurs, sur des multiples du fondamental, de façon à donner l'illusion d'une source sonore unique. La voix monocorde obtenue est plus agréable, pour la majorité des auditeurs, que la voix pseudo-chuchotée fabriquée avec le premier dictionnaire; ceci du moins quand le niveau sonore est assez important.

b) Amélioration de certains phonatomes

La schématisation adoptée pour certains phonatomes s'est avérée erronée ou incomplète à la suite des tests d'intelligibilité dont il sera question plus loin. Ces éléments ont été retouchés en fonction des confusions constatées, et de sonagrammes de voix différentes.

Il faut noter que la forme d'un phonatome peut changer légèrement en fonction de son entourage phonétique. Ainsi l'élément /tr/ dans le mot "citrouille" /sitruy/ et le même élément dans le mot "bijouterie" /bizutri/ ne sont-ils pas exactement superposables. La mise au point d'un élément /tr/ qui convienne dans tous les entourages phonétiques est une opération longue et délicate.

V.2.2.3. Les tests d'intelligibilité

Deux séries de tests ont été effectuées afin de chiffrer le taux d'intelligibilité de la voix synthétique. La première série (11.3.70) utilisait le dictionnaire voisé, mais non retouché; la seconde (24.11.70) bénéficiait des améliorations évoquées plus haut. Les mots utilisés, l'appareillage et l'auditoire étaient sensiblement identiques dans les deux cas.

a) Listes de mots

Le CNET nous a fourni les listes de mots utilisés lors des essais téléphonométriques (mots courants), ainsi que quelques phrases. Lors de la première épreuve les auditeurs ont pris en note 10 listes de 10 mots, contre 8 listes seulement dans la seconde; mais dans ce dernier cas il était demandé de noter également 5 phrases et 20 mots aléatoires. Dans les deux cas, l'épreuve proprement dite était précédée d'un petit texte d'introduction dont les sujets avaient la version écrite sous les yeux, afin de les familiariser avec le timbre quelque peu insolite de la voix synthétique.

b) Appareillage

Les deux séries de mots ont été synthétisées au moyen de l'Icophone III, et enregistrées sur magnétophone professionnel (Nagra). Elles ont été diffusées dans la même salle au moyen de chaînes d'écoute pratiquement identiques.

...../

c) Auditoire

L'auditoire était dans les deux cas constitué par le personnel du Laboratoire de Mécanique Physique de St-Cyr-l'Ecole. La plupart des régions linguistiques françaises étaient représentées; l'âge des sujets variait entre 20 et 60 ans, et l'on trouvait des professions aussi diverses que possible : mécaniciens, ajusteurs, secrétaires, techniciens, assistants, professeurs et étudiants.

A la première série ont participé 28 auditeurs, contre 24 à la seconde. Dix-neuf personnes ont participé aux deux séries; les personnes n'ayant subi le test que la seconde fois ont donné des résultats tout-à-fait comparables aux autres, ce qui permet d'exclure l'éventualité d'une mémorisation des mots, d'ailleurs bien improbable à 8 mois d'intervalle.

d) Résultats

Les mots notés par les auditeurs n'ont été comptés comme corrects que s'ils étaient parfaitement compris. Les résultats sont donnés sous forme d'histogrammes (fig. V-11) qui mettent en évidence les progrès réalisés. En moyenne le taux d'intelligibilité est passé de 65 à 88 %. La dispersion des résultats a fortement diminué.

La seconde série de tests comprenait 20 mots aléatoires de 5 phonatomes (6 phonèmes). La correction de ces mots phonème par phonème a permis de constater une bonne coïncidence avec les résultats établis sur les mots CNET (80 % de "phonèmes" correctement perçus). Les mots aléatoires peuvent donc être considérés comme des logatomes, avec le même défaut (les sujets sont gênés pour les transcrire en français) et la même qualité (absence de signification) ; ils présentent cependant l'avantage de mieux représenter la langue (rappelons que les phonatomes normalisés sont en espéranto ! ...) et donc à validité égale de nécessiter de moins longues listes de mots.

La qualité des phrases n'a été testée que lors de la seconde série, et a fourni un taux de 98 %. Ce chiffre est en parfait accord avec le précédent, si l'on se réfère aux courbes de KRYTER, classiques en téléphonométrie (bib. II-6). On peut donc déduire de cette étude que l'intelligibilité de la voix synthétique est d'ores et déjà supérieure à celle du téléphone.

Il convient de remarquer, à la suite de A. MOLES, que la parole synthétique, toujours semblable à elle-même, deviendra dans un proche avenir plus intelligible que la parole humaine, de même qu'un texte imprimé est plus lisible qu'un texte manuscrit car les caractères en sont normalisés.

Cette particularité trouve d'ailleurs une application en psycho-physiologie : la parole synthétique (speech-like sounds) constitue un stimulus parfaitement calibré et reproductible pour les tests de perception.

V.2.3. L'Icophone IV et l'anamorphose

La mise au point du dictionnaire sur l'Icophone III a permis de donner à la voix synthétique une intelligibilité quasi-totale. Pour passer de cette voix

..../

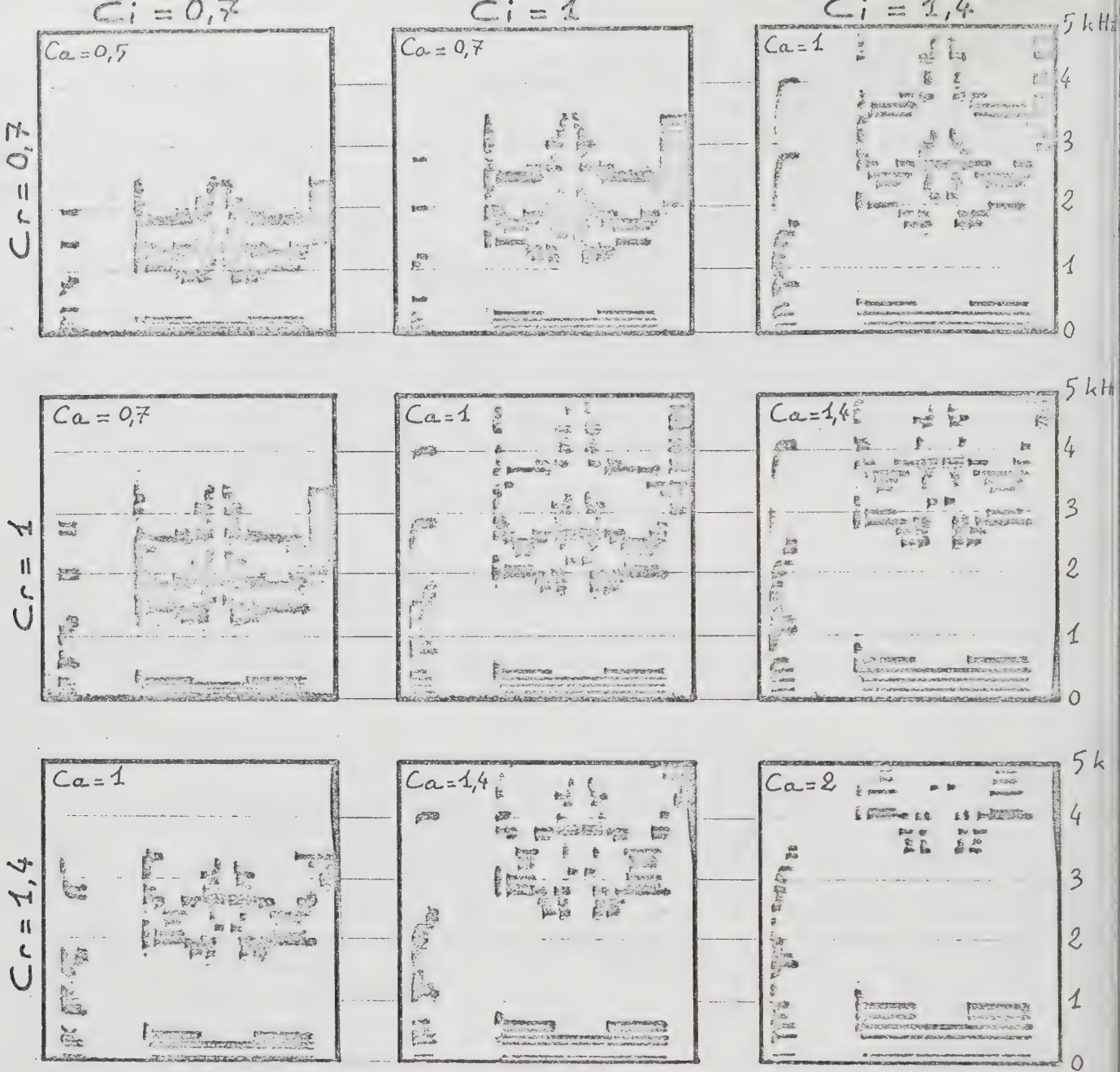


Fig V-13 - Diverses anamorphoses de la séquence "ATUVUS"..., synthétisée par l'Icophone IV, à coefficients d'intonation Ci et d'anamorphose relative Cr variables. Le coefficient d'anamorphose absolue Ca est indiqué dans chaque cas.

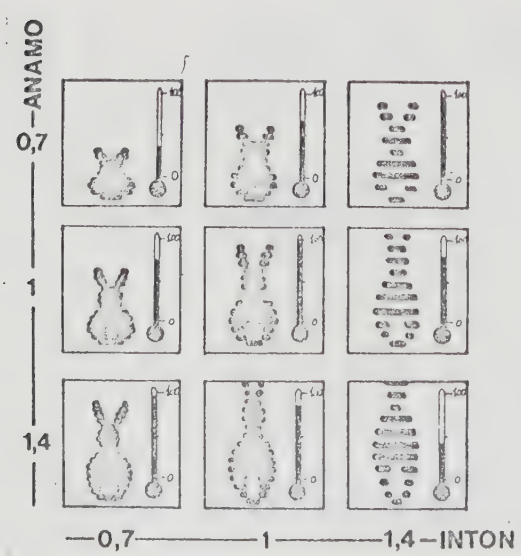


Fig V-14 - Analogie graphique avec l'expérience précédente. La température indiquée par le thermomètre représente l'intelligibilité de la voix synthétique (sur des mots aléatoires), ramenée à 100 pour le cas Ca = Cr = Ci = 1.

"utilitaire" à une voix plus élaborée, il convenait de pouvoir lui donner une intonation. C'est dans ce but qu'a été conçu l'Icophone IV, dont les 44 oscillateurs peuvent être modulés en fréquence tout en restant harmoniques (bib. V-12, V-13, IV-12). Mais l'Icophone IV, commandé par le même calculateur que l'Icophone III, permet aussi de pratiquer une anamorphose fréquentielle (bib. V-14). Cette particularité nous a permis d'étudier quantitativement l'intelligibilité de la voix anamorphosée.

V.2.3.1. Coefficients d'anamorphose et coefficient d'intonation

Il faut bien distinguer deux types de hauteurs : la hauteur de la voix, d'une part, c'est-à-dire la fréquence du fondamental ou l'écartement des raies du spectre, et la hauteur du squelette sémantique, d'autre part, qui s'apparente à une composante du timbre et ne peut être que repérée par rapport à une voix de référence.

Si nous prenons pour référence la voix de l'Icophone III, avec un fondamental de 100 Hz et le dictionnaire actuel, parfaitement défini et reproductible, nous pouvons définir sur l'Icophone IV un "coefficient d'intonation", C_i, rapport du fondamental à la fréquence 100 Hz, et deux coefficients d'anamorphose. Le premier, coefficient d'anamorphose relatif C_r, est le coefficient par lequel sont multipliées les fréquences du squelette sémantique lorsque le fondamental reste à 100 Hz. C'est ce coefficient qui est pris en compte par le programme ANAMO. Le second, coefficient d'anamorphose absolu Ca, est le produit des précédent :

$$Ca = C_i \cdot C_r$$

V.2.3.2. Résultats

La figure V-13 présente le résultat acoustique du jeu des coefficients C_i et C_r, qui prennent chacun les trois valeurs 0,7; 1 et 1,4. La séquence représentée est le début de la phrase synthétique "As-tu vu ce fameux lapin". Le long de la diagonale Sud-Ouest - Nord-Est, le coefficient C_a reste égal à 1 : le squelette sémantique est inchangé par rapport à la voix de référence, seule change la hauteur du spectre de raies. Celle-ci varie d'une octave le long de l'autre diagonale (N.O - S.E.), et le coefficient absolu passe de 0,5 à 2, ce qui représente une variation très grande (2 octaves) par rapport aux extrêmes rencontrés dans la voix réelle (1 octave). Horizontalement C_i varie seul, C_r étant fixé, et vice-versa dans la direction verticale.

L'expérience est schématisée selon une analogie visuelle dans la figure V-14, disposée comme la précédente. On a, en outre, indiqué dans chaque cas par la hauteur d'un petit thermomètre le résultat entre 0 et 100 de tests d'intelligibilité effectués avec la voix anamorphosée. Ces tests ont été effectués avec des mots aléatoires notés phonétiquement par 7 auditeurs. On a pris comme base le chiffre 100 dans le cas $Ca = C_r = C_i = 1$ (case centrale de la figure), bien que l'intelligibilité fût alors de 80 %. Mais ce taux, estimé sur des mots aléatoire, correspond à une valeur plus élevée sur des mots ou phrases courants.

On peut constater que le taux de compréhension varie peu (de l'ordre de 20 %) tant que l'anamorphose absolue reste dans les limites de 0,7 à 1,4 qui correspondent sensiblement aux voix réelles, et qu'elle n'est pratiquement pas fonction du timbre de la voix. Il est d'ailleurs vraisemblable que .../

la perte observée dans les cas extrêmes provient plus de la mutilation du squelette sémantique que d'une anamorphose excessive. En effet, dans les anamorphoses de coefficient relatif inférieur à 1 une partie de l'information est perdue, car comprimer le squelette sémantique revient à diminuer la finesse de résolution fréquentielle. Dans les anamorphoses de coefficient relatif supérieur à 1 les fréquences élevées sont perdues : une anamorphose de coefficient relatif $C_r = 1,5$ par exemple opère la transposition de la fréquence n° 30 vers la fréquence n° 45, qui n'est pas prévue sur l'Icophone IV; ces anamorphoses s'accompagnent donc d'un effet de filtrage passe-bas.

V.2.4. Comparaison d'un phonatome à tous les autres

Les phonatomes d'un même dictionnaire étant normalisés en temps et en fréquence, il est intéressant de les comparer entre eux, afin de détecter d'éventuelles confusions. Ce contrôle permet une meilleure mise au point des phonatomes synthétiques : chaque phonatome doit, tout en restant intelligible, se différencier au maximum de ses voisins. Deux difficultés se présentent :

- D'une part, le critère de comparaison utilisé doit être pertinent, c'est-à-dire avoir quelque rapport avec l'appréciation auditive. Le critère COMPA (taux de coïncidence, voir annexe II) nous a donné dans ce domaine les meilleurs résultats.

- D'autre part les phonatomes doivent être complètement normalisés en temps, et pas seulement en durée totale. Pour comparer valablement /pa/ et /pœ/, par exemple, l'explosion du /p/ doit se trouver au même endroit dans les deux cas. En d'autres termes, il faut opérer un remaniement préalable du dictionnaire et placer le pivôt (instant où le squelette sémantique est le plus instable) exactement au milieu du phonatome. Un tel programme a été écrit par A. CALINET.

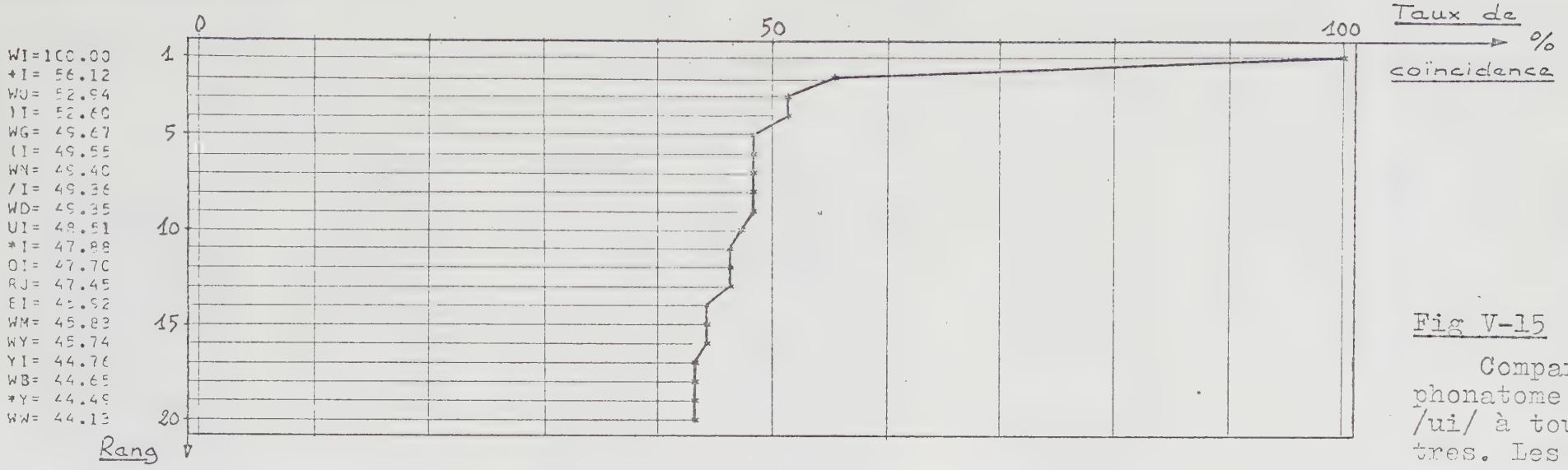
La fig. V-15 présente un exemple des résultats obtenus à partir du dictionnaire de Mlle CASTELLENGO, adapté en vue de la reconnaissance, c'est-à-dire recadré en 5 événements de 32 fréquences, les valeurs spectrales étant étendues à l'intervalle 0-9. Le phonatome /ui/ (mot "oui"), comparé à ses voisins, présente une émergence très satisfaisante, puisque le plus proche, /oi/, donne lieu à un taux de coïncidence de 56 %. De plus, la succession des plus proches voisins est logique sur le plan acoustique, comme on peut le vérifier d'après leur représentation temps-fréquence.

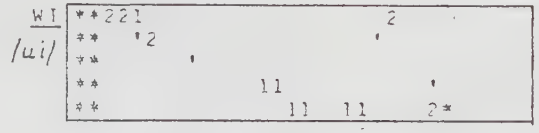
L'étude de la série des plus proches voisins nous a permis, dans d'autres cas, de détecter certaines erreurs : l'apparition du phonatome /rj/ comme premier voisin de /la/, par exemple, correspond évidemment à une erreur dans le dictionnaire (erreur dûe non pas à la schématisation, mais à de malencontreuses manipulations de fichiers).

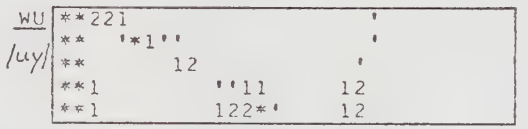
Naturellement il ne s'agit ici que d'illustrer une méthode, sans donner des résultats quantitatifs de portée générale : il n'est pas très légitime de ramener les phonatomes à 5 événements seulement, du moins pour la synthèse, et le pivôt n'est pas toujours bien replacé sur le 3° événement. Il est cependant souhaitable d'approfondir l'étude et l'utilisation de cette méthode, et de mettre les résultats obtenus en corrélation avec, d'une part, l'information associée à chaque phonatome, et d'autre part les tests d'intelligibilité effectués sur la voix synthétique. Il est vraisemblable, en effet, que plus un phonatome est acoustiquement original, plus son émergence par rapport aux plus proches

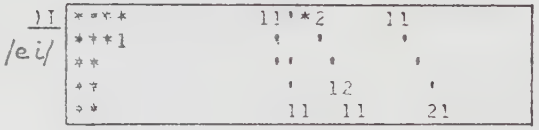
..../

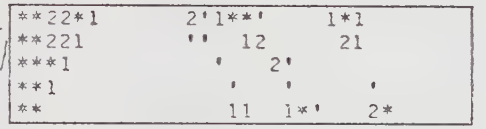
PHONATOME DE REFERENCE WI /ui/

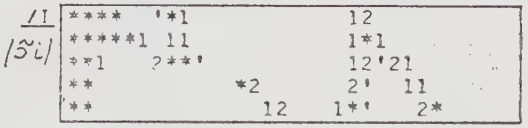


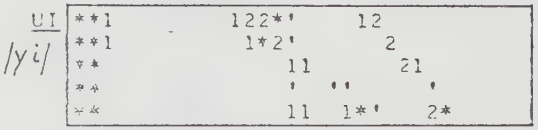


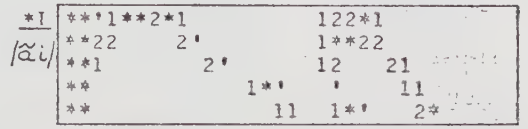












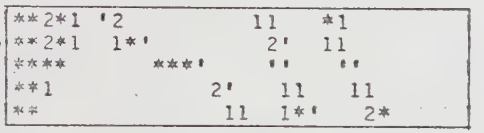


Fig V-15

Comparaison du phonatome normalisé /ui/ à tous les autres. Les phonatomes du dictionnaire sont classés par ressemblance décroissante. Les premières valeurs de la courbe caractérisent l'originalité du phonatome de référence. Dans cet exemple l'émergence de /ui/ est excellente, et la succession des phonatomes voisins est logique : on trouve toujours une voyelle commune avec le phonatome de référence.

voisins devrait être grande, et plus sa fréquence d'occurrence dans le langage devrait être faible; réciproquement, les confusions décelées dans le dictionnaire devraient apparaître dans les réponses des auditeurs.

V.3. LES APPLICATIONS DE LA SYNTHESE

Si l'on met de côté l'intérêt purement scientifique de la synthèse - meilleure connaissance des structures acoustiques et phonétiques de la parole - on peut distinguer deux types d'applications susceptibles d'un développement important dans les années à venir.

V.3.1. La réponse vocale automatique

Un synthétiseur transformant une séquence de mots orthographiés ou phonétiques en message vocal, peut être considéré comme un périphérique de calculateur, ou unité de réponse vocale.

V.3.1.1. Les unités de synthèse classiques

Actuellement les unités de réponse vocale que l'on trouve sur le marché comprennent un certain nombre de mots enregistrés sur un tambour sous forme analogique (procédé magnétique ou optique). Ces mots sont disponibles immédiatement sur ordre du calculateur. Il ne s'agit pas à proprement parler de synthèse, mais d'enregistrement et lecture par mots (type horloge parlante).

Une amélioration importante consiste à stocker les mots sous forme numérique, après réduction d'information par l'intermédiaire d'un vocoder. L'utilisation du vocoder est ici pleinement valable, puisque l'on ne recherche pas une économie maximum, et que les données enregistrées peuvent être retouchées manuellement jusqu'à fournir une synthèse parfaitement intelligible. C'est la voie suivie par I.B.M. et par le C.N.E.T. (bib. V-6).

Malgré cela, la limitation du vocabulaire à quelques centaines de mots constitue une entrave au développement des unités de réponse vocale; encore pour stocker de tels vocabulaires est-il nécessaire de disposer d'un calculateur de grande taille. L'apparition d'unités pratiquement autonomes, à vocabulaire illimité et réponse immédiate, est susceptible de bouleverser le marché.

V.3.1.2. Applications potentielles des unités de réponse vocale.

a) Synthèse vocale de messages élaborés par un calculateur : gestion des stocks, réservation de places, agences de presse, bourse, télégrammes téléphonés, gares, centres de calcul, etc... Tout système automatisé dont le fonctionnement doit être suivi par plusieurs personnes ayant une activité annexe bénéficierait de cette application pour deux raisons :

..../

- les personnes concernées ne seraient plus rivées à l'écran cathodique ou à la machine télétype.

- le coût d'une extension géographique est minime (coût d'un haut-parleur 10 F; coût d'un télétype 10 000 F., coût d'un écran 20 à 30 000 F).

Un aspect non négligeable de cette application serait la possibilité offerte aux aveugles de s'adapter facilement à toutes les activités utilisant l'informatique.

b) prothèse pour muets

Un synthétiseur est une machine parlante; miniaturisée et adaptée à une utilisation manuelle, elle pourrait trouver dans ce domaine un champ d'application insoupçonné à présent.

D'application plus lointaine mais nullement utopique, la "machine à lire à voix haute" présenterait également un intérêt certain pour les aveugles (il existe en France et à l'étranger plusieurs systèmes de lecture automatique des caractères imprimés).

V.3.1.3. L'Icophone comme unité de réponse vocale

Tous ces débouchés font que divers constructeurs cherchent activement à mettre au point des unités à vocabulaire illimité, réponse immédiate et faible coût. L'Icophone est actuellement développé dans ce sens (Icophone V, en cours de mise au point avec l'aide de la Délégation à l'Informatique). L'Icophone se prête particulièrement bien à ce genre d'application, son vocabulaire est par définition illimité, puisque la commande est phonétique. Notons d'ailleurs que le débit requis sur la ligne de commande pour une synthèse en temps réel est d'environ 50 bits/seconde, c'est-à-dire le débit de l'information sémantique elle-même (rappelons que le débit requis pour commander un télétype est de 110 bits/seconde). La réponse est immédiate, ou plus exactement le délai nécessaire entre la réception du message phonétique et l'émission du message vocal dépend de la logique et du programme utilisés : l'Icophone III commandé par le 1130 requiert un délai de l'ordre de 1 seconde pour 70 caractères phonétiques.

V.3.2. La transmission de la parole à très faible bande

Le débit de commande d'une unité telle que l'Icophone ne dépasse guère 50 bits/seconde, soit environ 40 fois moins que les synthétiseurs du type VOCODER les plus économiques. Mais pour envisager une application à la transmission téléphonique, il est nécessaire de disposer du système de reconnaissance automatique, transformant la parole réelle en suite phonétique. Nous verrons au chapitre suivant ce que l'on peut penser de ce problème. Remarquons cependant dès à présent que la mise au point d'un système de reconnaissance "par mots" opérant sur la parole continue ouvrirait dans ce domaine des perspectives révolutionnaires : la limitation du vocabulaire à $2^{10} = 1024$ mots, par exemple, impliquerait pour un débit d'environ 2 mots par seconde, la transmission de 2 x 10 = 20 bits/seconde seulement.

..../

V.4. CONCLUSION

Avec la synthèse de la parole, nous sommes entrés dans un domaine très concret, puisqu'il a même été question de prix de revient. Nous espérons pourtant avoir montré que la réussite de l'Icophone résultait directement de l'ensemble des idées exposées aux chapitres III et IV, idées qui se trouvent ainsi confirmées dans les faits. Nous abordons à présent avec la même doctrine le domaine encore relativement vierge de la reconnaissance automatique de la parole.

VI - LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

La reconnaissance automatique de la parole est un problème récent (les premières réalisations datent de 1951), lié au développement de l'automatisme et de l'informatique. La recherche est motivée dans ce domaine par les nombreuses applications possibles : commande vocale de machines (machines-outils, tri postal, etc..), machine à écrire phonétique, dialogue avec les calculateurs, etc... Cependant, après un élan considérable d'une douzaine d'années, la recherche a marqué le pas : une certaine déception s'est manifestée parmi les équipes les plus avancées (voir par exemple N. LINDGREN, bib. VI-1, ou J.L. FLANAGAN, bib. II-6 et VI-2). Cette attitude nous amène à nous demander si le problème peut accepter une solution.

Tout dépend des hypothèses et des limites assignées à la recherche. Si l'on attend d'un système de reconnaissance qu'il effectue des opérations aussi raffinées que ne le fait le cerveau humain, qui parvient à saisir la signification de messages informes, mal articulés ou noyés dans le bruit, dans ce cas il vaut mieux attendre que la connaissance des mécanismes perceptifs soit plus avancée. Si, par contre, on se fixe un objectif limité, comme la reconnaissance de quelques dizaines de mots prononcés par un seul locuteur, alors le problème est soluble et même largement résolu. Encore les données du problème sont-elles souvent faussées par le choix des locuteurs, les contraintes de prononciation qui leur sont imposées, le choix des mots à reconnaître, l'adaptation de l'homme à la machine. Toutes ces considérations font qu'il est extrêmement difficile d'estimer à leur juste valeur les résultats fournis par un système de reconnaissance.

Pour notre part nous considérons un système de reconnaissance comme un modèle actif de la perception. Ce modèle est certes fragmentaire, pour l'instant du moins. Mais il fournit des résultats objectifs, concrétisés par des taux de reconnaissance. Il permet donc, si sa structure est suffisamment souple - et la simulation numérique assure cette souplesse - de tester objectivement la valeur fonctionnelle des hypothèses dégagées par l'analyse et la synthèse de la parole, ou par l'étude de la perception. Dans ce sens, le meilleur taux de reconnaissance obtenu est moins intéressant que les moyens mis en oeuvre pour le faire évoluer.

Cette optique n'exclut nullement la définition d'appareillages plus spécialisés; bien au contraire, elle évite une orientation trop rigide de la recherche. Nous verrons par exemple que les mêmes idées de base donnent lieu à deux versions différentes du programme, l'une dirigée vers la reconnaissance analytique, par phonatomes, et l'autre vers la reconnaissance globale de mots ou de phrases.

Avant de décrire ces expérimentations, il nous semble utile de présenter l'essentiel des méthodes et des moyens utilisés en reconnaissance vocale. Nous présenterons les principaux aspects de la reconnaissance des formes, en expliquant pourquoi les théories actuelles, qui se rapportent surtout à la phase de décision, ne sont guère applicables à la parole. Nous évoquerons ensuite l'important problème de l'apprentissage, puis nous examinerons les problèmes spécifiques de la reconnaissance vocale.

Nous décrirons enfin quelques réalisations-types, dont nous tenterons de dresser le bilan à la lumière de l'expérience acquise en analyse et synthèse de la parole.

...../

VI.1. METHODES ET MOYENS DE LA RECONNAISSANCE VOCALE

VI.1.1. Théorie de la reconnaissance des formes

VI.1.1.1. Position du problème

La reconnaissance des formes est une discipline mathématique qui naquit voici une vingtaine d'années. Elle concernait surtout, à l'origine, la reconnaissance des formes visuelles, et tout particulièrement la reconnaissance des caractères de l'écriture. Le terme de forme n'a pas exactement le sens de totalité perceptive que nous lui avons prêté au chapitre III, dans le cadre de la Gestalttheorie : une forme est ici une certaine description d'un objet choisi dans un ensemble partitionné en familles, selon des critères plus ou moins arbitraires. Reconnaître une forme, c'est décider qu'elle représente un objet appartenant à une famille donnée. Pour éviter toute confusion de termes, nous emploierons plutôt les mots "image" s'il s'agit de formes graphiques, "configuration" ou "occurrence" dans les autres cas.

Soit par exemple un ensemble d'objets plans comprenant trois familles : les objets ronds, les objets rectangulaires, et les objets triangulaires. La figure VI-1 montre les images de quelques-uns de ces objets, obtenues au moyen d'un certain système de description : les images proviennent ici d'une projection des objets dans un plan parallèle à leur support; seul le contour est conservé à l'exclusion des autres caractéristiques telles que couleur, masse, matière, odeur, etc... La figure VI-2 montre diverses images du même objet, qui se déduisent les unes des autres par des transformations géométriques simples : translation, rotation, symétrie. Connaissant les critères de classification des objets, il s'agit de décider pour chaque image si elle provient de telle ou telle famille d'objets. L'ensemble infini des images possibles comporte trois classes d'équivalence.

La reconnaissance, évidente pour un cerveau humain, est assez facile à effectuer mathématiquement sur cet exemple simple : il suffit de détecter le nombre de points anguleux de chaque contour. Encore ce principe présente-t-il déjà certaines difficultés de mise en oeuvre, variant selon la présentation matérielle des images : tracés, équations, tableaux de nombres, etc...

Les choses se compliquent si l'on introduit une nouvelle famille d'objets, par exemple les objets trapézoïdaux : il faut alors un critère de classification supplémentaire (parallélisme des côtés opposés, dans les images à 4 angles). Si l'on considère aussi les losanges, il faut envisager la présence d'un angle droit dans les parallélogrammes, etc... On voit s'esquisser un problème de stratégie propre à chaque ensemble d'objets considéré, avec ses critères de classification : faut-il détecter ici d'abord le parallélisme, ou d'abord les angles droits ? La stratégie sera plus complexe si l'on considère d'autres types d'objets, comme ceux de la figure VI-3 : hexagone régulier, ellipse, ou objets plus difficiles à décrire géométriquement.

Nous n'avons considéré jusqu'à présent que des images définies sans ambiguïté. Mais en général les images sont, soit floues (ne serait-ce qu'à cause de la quantification), soit légèrement déformées par rapport aux images-types de chaque classe d'équivalence. La fig. VI-4 représente quelques-unes de ces déformations, qui nous replongent dans la Gestalt-Theorie. La notion d'identification entre une image donnée et une image-type doit être remplacée par la notion de parenté, ou de voisinage, objectivée par une grandeur appelée distance : l'image donnée est reconnue comme appartenant à la classe de l'ima-

...../

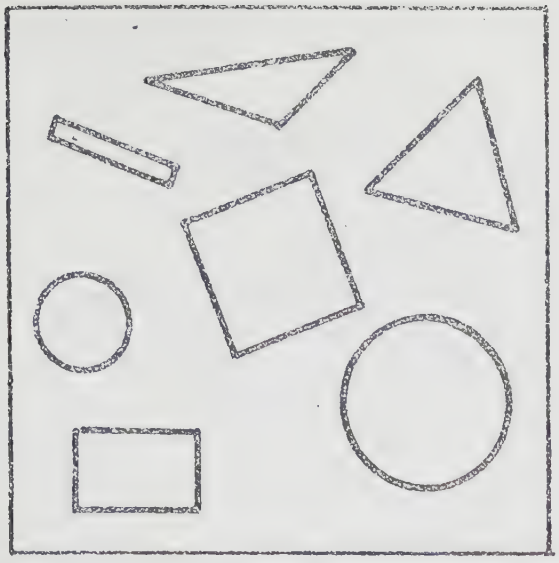


Fig VI-1

Images d'objets plans, appartenant à 3 familles aisément paramétrisables : objets ronds, triangulaires, rectangulaires.

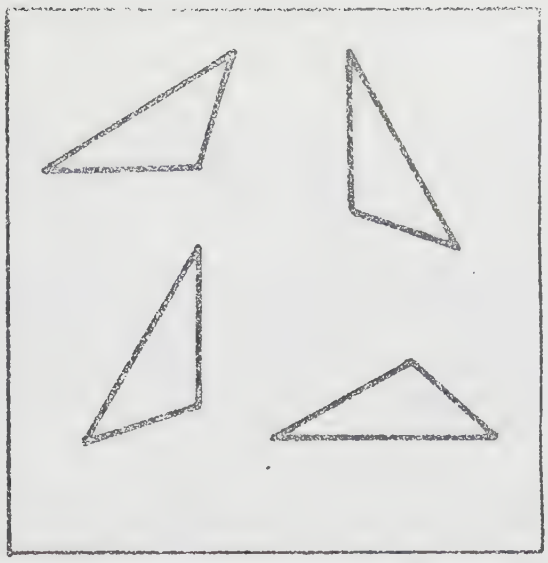


Fig VI-2

Diverses images du même objet, se correspondant dans des transformations géométriques simples : translation, rotation, symétrie plane.

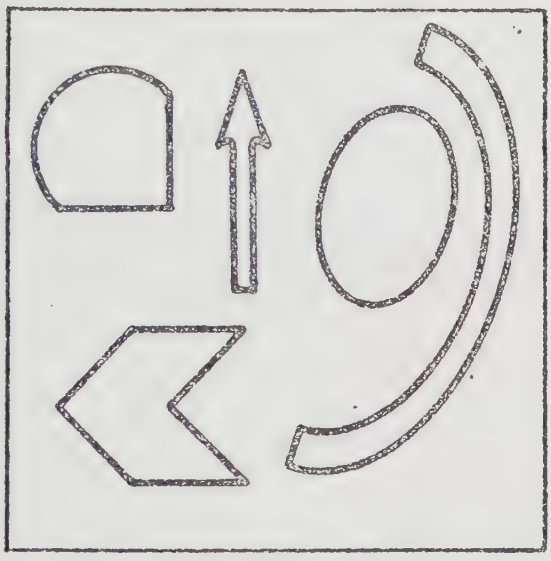


Fig VI-3

Quelques images d'objets dont la représentation paramétrique est moins aisée.

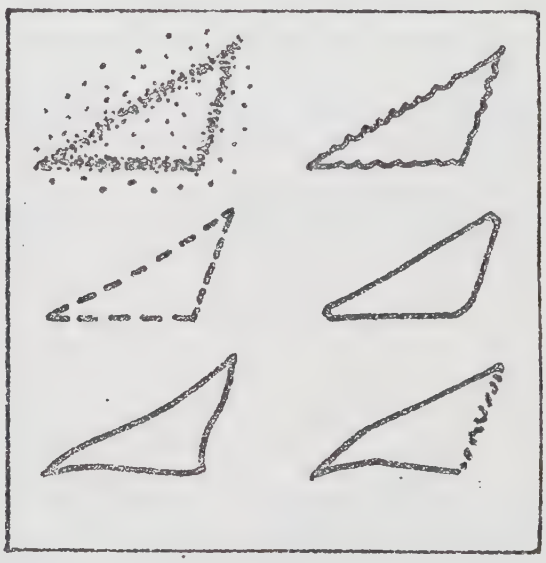
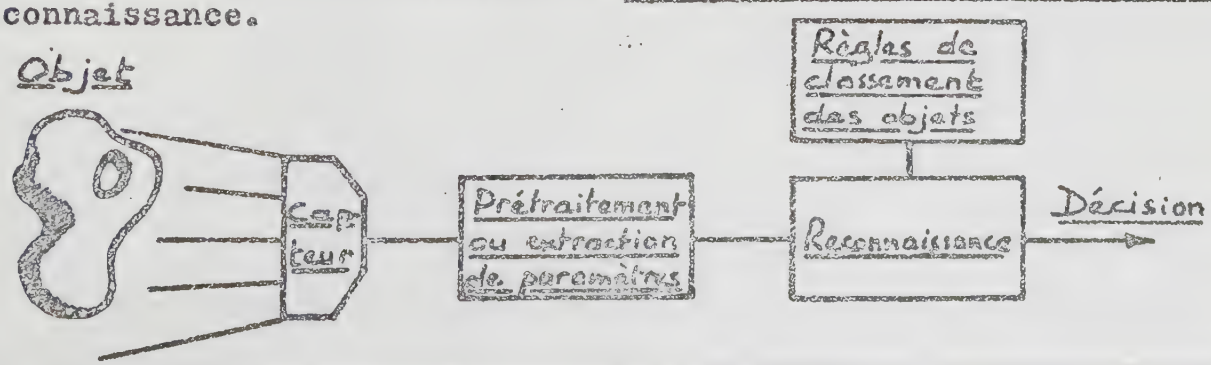


Fig VI-4

Quelques cas dans lesquels la représentation paramétrique est inadéquate : il vaut mieux utiliser une méthode de corrélation (qui implique normalisation et recadrage).

Fig VI-5 (ci-dessous)

Organisation d'un système de reconnaissance.



ge-type la plus voisine. Les résultats sont évidemment fonction du type de déformation et de la définition mathématique adoptée pour la distance.

Dans les problèmes réels une complication importante intervient lorsque les images à identifier sont imbriquées les unes dans les autres, ou simplement juxtaposées : c'est le cas de l'écriture manuscrite, de la parole, de tous objets photographiés sur un fond hétérogène, etc... La séparation des images à reconnaître est un problème ardu, que le cerveau résout avec une facilité dérisoire par rapport aux machines actuelles.

La reconnaissance est donc une opération passablement arbitraire. Elle suppose d'emblée que les règles de classification des objets peuvent être traduites par des classes d'équivalence des images, et que les images contiennent toute l'information nécessaire à leur séparation et à leur "reconnaissance".

VI.1.1.2. Organisation d'un système de reconnaissance

Tout système de reconnaissance se compose de trois organes distincts (fig. VI-5) :

- Un capteur, qui fournit les images à reconnaître en fonction des objets
- Un organe de prétraitement, qui selon le cas effectuera les cadrages , lissage, normalisations nécessaires, ou l'extraction de paramètres.
- Un organe de décision, dans lequel sont stockés de façon explicite ou implicite les critères de classification des objets, et les règles de parenté des images déformées.

VI.1.1.3. Théorie de la reconnaissance

A vrai dire, il n'existe pas encore de théorie unifiée de la reconnaissance des formes. Seule la dernière phase a fait l'objet d'études suffisamment générales, qui visent à déterminer les meilleures méthodes de décision (bib. II-3, III-6, VI-3). Ces méthodes se divisent en deux familles, selon la structure des classes d'équivalence.

a) Méthodes de corrélation

Ces méthodes s'appliquent lorsque les configurations d'une même classe d'équivalence se correspondent dans des transformations géométriques, ou tout au moins mécanisables. On définit alors, selon une même norme, autant de configurations-types que de classes d'équivalence. La configuration à reconnaître est d'abord isolée et normalisée, puis comparée à toutes les configurations-types, et on choisit la configuration ayant donné lieu à la meilleure corrélation.

b) Méthodes paramétriques

Lorsque les configurations à reconnaître peuvent être décrites par un ensemble de N paramètres, on représente chaque configuration par un point dans un hyper-espace à N dimensions. Les classes d'équivalence sont alors

...../

représentées par des ensembles de points dans cet espace. Si les paramètres sont bien choisis, ces ensembles de points sont groupés dans des domaines bien délimités. Reconnaître une configuration donnée revient à établir l'appartenance du point représentatif à l'un de ces domaines. Il s'agit alors d'un problème mathématique largement étudié, dont la résolution varie selon la forme des domaines, leur interpénétration éventuelle, la distribution des points, le coût des erreurs, etc... On peut distinguer deux techniques différentes :

- Séparation des domaines par des hypersurfaces; dans le cas le plus simple on utilise des hyper-plans, mais la séparation obtenue ne peut être universelle (la fig. VI-6 donne un équivalent en deux dimensions).

- Définition de points-références, par exemple les barycentres des divers domaines, et recherche de la distance minimum du point donné à ces points-références (fig. VI-7). La notion de distance est aménagée en fonction des besoins et des moyens de résolution : distance absolue, distance quadratique ou de degré supérieur, distance de HAMMING (nombre d'éléments différents entre deux nombres binaires) etc...

Nous ne développerons pas l'aspect mathématique qui nous entraînerait trop loin du sujet. Nous remarquerons seulement que l'approche mathématique est difficilement applicable à la reconnaissance de la parole, dont les problèmes se situent actuellement à un tout autre niveau, celui du prétraitement et de la définition même des objets à reconnaître. Avant d'aborder les problèmes spécifiques de la reconnaissance vocale, nous dirons quelques mots de l'apprentissage.

VI.1.2. Les problèmes d'apprentissage

L'apprentissage consiste à introduire dans le système de reconnaissance les règles de classement des objets et de parenté des images par un processus d'essais, d'erreurs et de corrections, au lieu de les imposer de manière arbitraire et figée. En général l'apprentissage dispense de la description exhaustive et explicite de ces règles ou références.

L'apprentissage procède généralement par comparaison entre la décision prise par le programme de reconnaissance et la décision prise par un "opérateur idéal", c'est-à-dire pratiquement par l'expérimentateur (notons que, comme en synthèse, l'expérimentateur est souvent très mauvais juge). Certaines études visent à optimiser ce processus (bib. VI-4). Cependant l'apprentissage se trouve matérialisé dans deux systèmes qui ont fait surgir de grands espoirs en reconnaissance des formes, le perceptron et la matrice d'apprentissage.

VI.1.2.1. Principe du perceptron (bib. II-13, VI-5)

Présenté en 1960 par E. ROSENBLATT, le Perceptron est un système capable d'"apprendre" une configuration, et ensuite de l'identifier si elle lui est présentée à nouveau dans les mêmes conditions. Il comprend trois couches d'éléments (fig. VI-8) : les éléments sensoriels quantifient la configuration étudiée et la transforment en un ensemble de grandeurs mesurables; les N éléments de la couche associative sont connectés aléatoirement à plusieurs élé-

.....

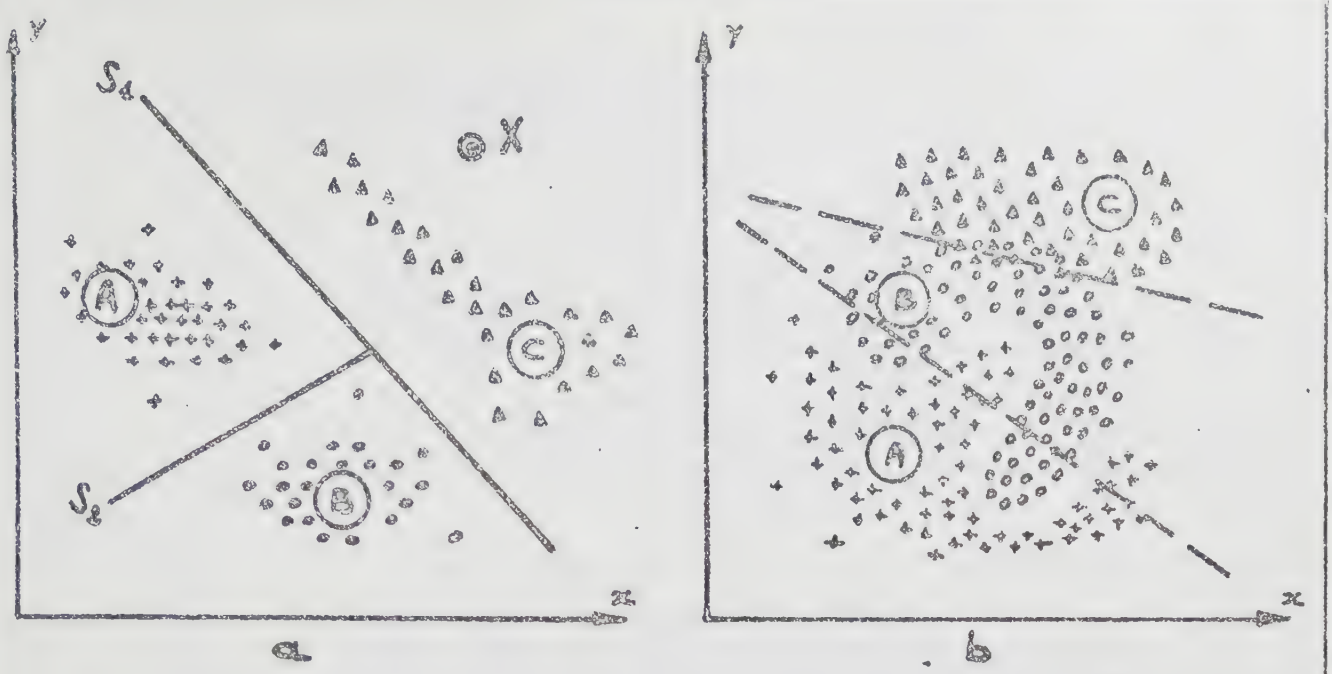


Fig VI-6

Séparation linéaire, dans un espace à deux dimensions.
a) Cas où une séparation linéaire est possible. Les nuages A, B et C sont compacts et bien distincts. La configuration X est "reconnue" comme appartenant à la famille C.
b) Cas où une séparation linéaire est impossible ou très difficile à mettre en oeuvre. Les familles A et B sont imbriquées l'une dans l'autre : une séparation par des segments de droite serait très complexe. Les familles B et C s'interpénètrent légèrement : les paramètres x et y sont sans doute mal choisis.

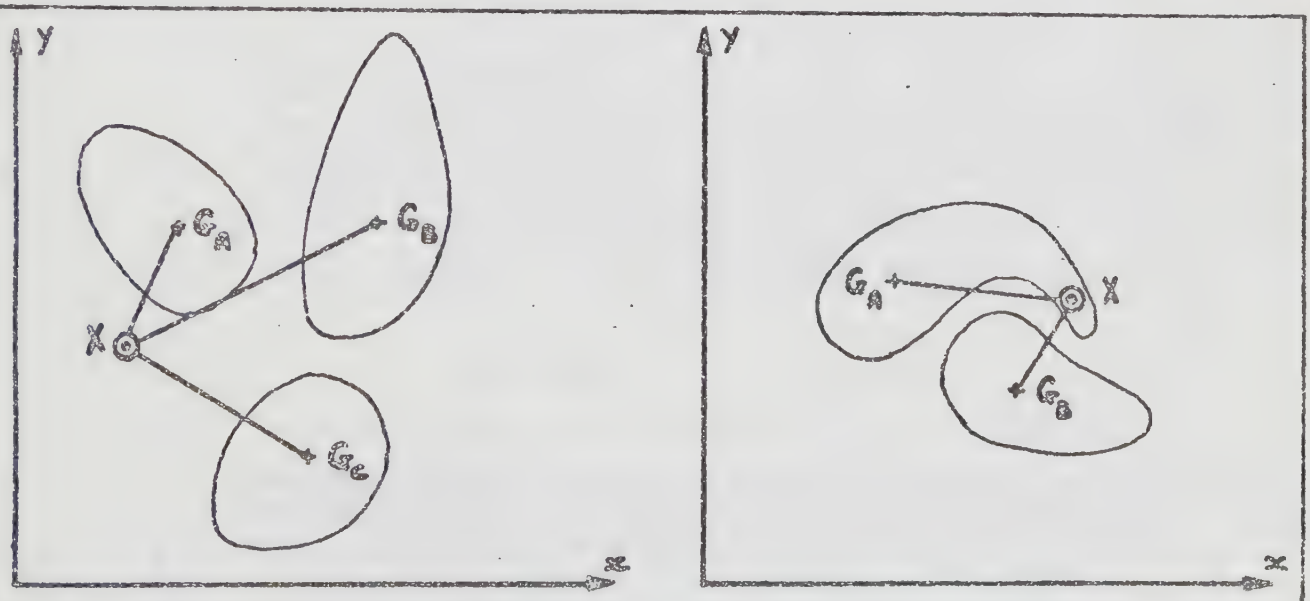


Fig VI-7

Identification d'une configuration X par recherche de distance minimale. Les nuages A, B et C sont représentés par leurs barycentres G_A, G_B et G_C. Cette méthode permet un classement par distance décroissante.
a) la configuration X est identifiée à l'objet A.
b) bien que faisant partie de la famille A, la configuration X est identifiée à l'objet B.

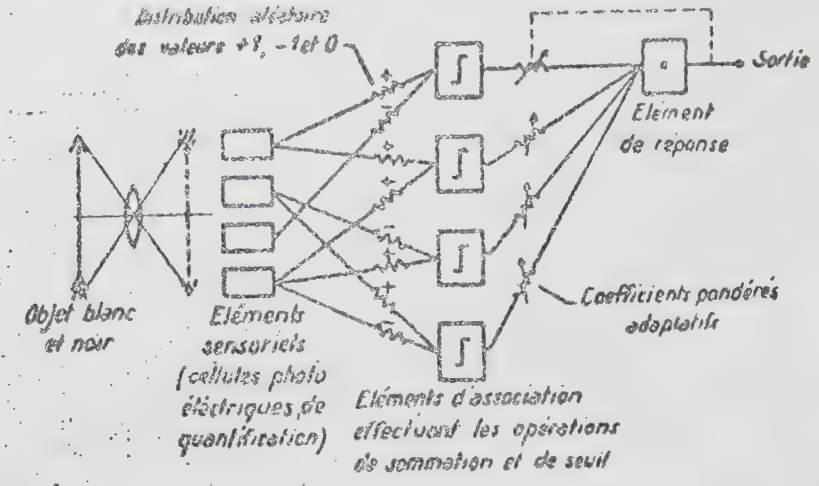


Fig VI-8

Principe du perceptron. A une classe d'objets est associée par apprentissage une sortie du perceptron. Mais l'apprentissage est très long et les possibilités de discrimination sont limitées (d'après J.S.BRYAN, cité par A.DEWEZE).

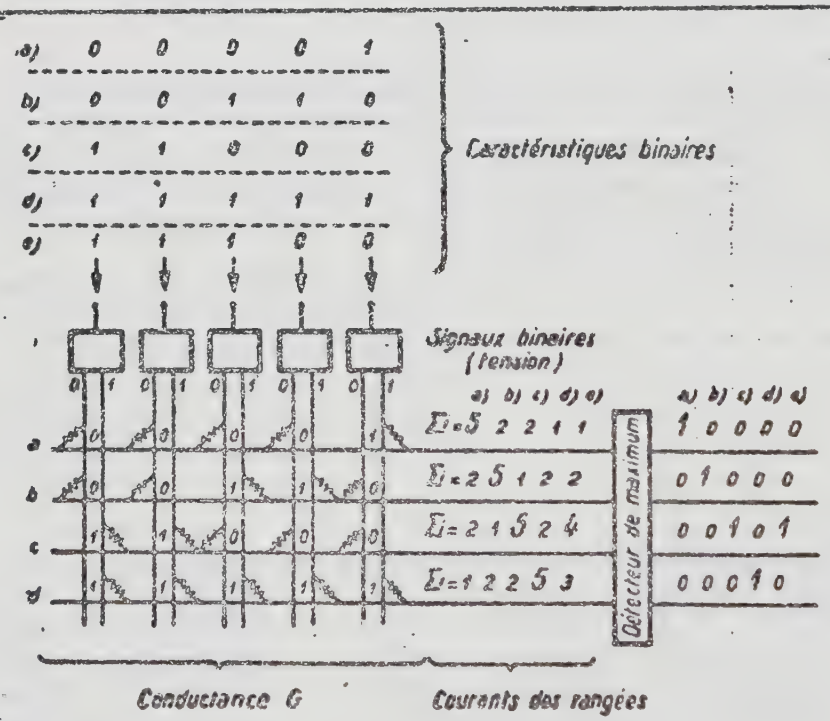


Fig VI-9

Principe de la matrice d'apprentissage. Dans les lignes de la matrice sont stockées les configurations binaires a(00001) , b(00110), c(11000), et d(11111). Si l'on présente ces mêmes configurations sur les colonnes de la matrice, on recueille sur chaque ligne une tension représentant la proximité avec chacune des références. La configuration supplémentaire e(11100) donne lieu à une corrélation maxima avec la référence c(11000).(d'après STEINBUCH et et PISKE, cités par A.DEWEZE).

ments sensoriels, et font la somme a_j des grandeurs associées. Enfin, chaque élément de réponse effectue une somme pondérée des grandeurs a_j soit

$$r = \sum_{j=1}^{N} c_j a_j$$

et délivre un signal binaire si cette somme dépasse un certain seuil.

A toute configuration originale correspond un ensemble original de valeurs a_j. On "apprend" les configurations X et Y au perceptron en ajustant les coefficients c_j de façon à obtenir

$$r(X) = 1 \quad \text{et} \quad r(Y) = 0$$

Les N réglages possibles semblent surabondants; cependant, ils permettent de rendre la réponse invariante dans certaines déformations ou certains déplacements de la configuration : on présente une configuration X_1 de la même classe d'équivalence que X, et l'on renforce les réglages des c_j qui fournissent simultanément les réponses

$$r(X) = 1$$

$$r(X_1) = 1$$

On continue ainsi pour une configuration X_2, puis X_3, etc... et l'on fait de même pour la classe d'équivalence de Y. On peut ainsi discriminer deux nouvelles configurations X' et Y'.

Avec M éléments de réponse binaire r_k (k = 1,2, ... M) on peut discriminer 2^M classes de configurations différentes. L'apprentissage devient évidemment très laborieux car les coefficients c_{kj} à ajuster sont au nombre de M x N. Ainsi, tous les cas examinés lors de l'apprentissage sont intégralement mémorisés de façon implicite dans les valeurs des coefficients c_{kj}. Les éventuelles simplifications de structure des classes d'équivalence, provenant de transformations géométriques simples telles que translation ou rotation, ne sont pas prises en compte dans ce processus.

VI.1.2.2. Principe de la matrice d'apprentissage

Le principe de la matrice d'apprentissage développée en particulier par STEINBUCH en 1962, revient à la comparaison rationnelle d'une configuration inconnue X avec N configurations de référence R_i (i = 1, 2, ... N).

Soient x_j (j = 1, 2, ... M) les composantes de X fournies par un ensemble de M capteurs, et r_{ij} les composantes des références R_i. L'ensemble des références peut être noté sous la forme d'une matrice r_{ij} de N lignes et M colonnes dont chaque ligne représente les composantes de R_i.

On définit une certaine opération de comparaison entre deux configura-

...../

tions, et l'on forme la matrice résultant de la comparaison terme à terme de X et des R_i . Par exemple, si l'on utilise comme critère de comparaison la distance absolue entre les valeurs, on forme la matrice

$$\left\{ \left| r_{ij} - x_j \right| \right\}$$

Reconnaître la configuration X revient à déterminer la ligne i de la matrice pour laquelle somme S_i est minima :

$$S_i = \sum_{j=1}^{M} \left| r_{ij} - x_j \right|$$

$$X \Longrightarrow R_n \quad ; \quad n \text{ tel que } S_n = \min S_i \quad (i = 1,2, \ldots N)$$

Présentée de cette manière, la matrice d'apprentissage apparaît comme une illustration banale de la méthode générale de reconnaissance par corrélation. Mais l'intérêt de cette formulation est de se prêter facilement à une réalisation analogique : la configuration R_i est présentée sous forme d'un ensemble de M tensions analogiques V_{ij} sur les fils verticaux (fig. VI-9)

Le fil horizontal i est porté au potentiel V_o , alors que les autres sont maintenus à zéro. En chaque noeud de la matrice se trouve matérialisée et mémorisée une relation entre V_o et V_{ij} (rapport, différence, etc, selon le type de distance choisie). Dans certains cas on peut cumuler sur la même ligne l'information provenant de plusieurs configurations de la même classe d'équivalence R_i , R'_i , R''_i , etc... Lorsque l'apprentissage est terminé on présente la configuration X sur les M fils verticaux; on obtient sur les N fils horizontaux des tensions U_i proportionnelles à la distance (ou à la proximité, selon la relation choisie) de la configuration X aux configurations R_i . Une simple détection de minimum ou de maximum suffit à terminer la reconnaissance.

La matrice d'apprentissage possède d'intéressantes propriétés de classement et d'invariance par rapport à certaines transformations des configurations à reconnaître. L'apprentissage est infiniment plus simple à réaliser que dans le perceptron. Cependant, aucun de ces systèmes ne se prête à une application immédiate dans le domaine de la parole.

VI.1.2.3. Limites de l'apprentissage

Il est bien évident qu'aucun système, ni aucun cerveau, ne peut reconnaître sans avoir appris. Il faut bien disposer de références, sous une forme ou sous une autre. Mais la distinction s'impose entre l'apprentissage-mémorisation des références, et l'apprentissage-mise au point, sous la direction d'un moniteur ou opérateur idéal, qui corrige au fur et à mesure les réglages du système jusqu'à obtenir un taux de réussite suffisamment élevé.

L'apprentissage, au sens de la reconnaissance des formes, n'est que rarement utilisable pour d'autres classes de configurations que celles qui

...../

ont strictement été étudiées (sinon on entre dans le domaine encore quasi-vierge des systèmes auto-adaptatifs). Deux aspects pratiques de l'apprentissage méritent d'être soulignés :

a) les configurations utilisées pour l'apprentissage doivent être parfaitement distinctes des configurations utilisées en reconnaissance; faute de quoi les résultats sont biaisés. Il est même à peine légitime d'effectuer l'apprentissage et la reconnaissance sur des mots prononcés à la suite les uns des autres par le même locuteur, dans des conditions identiques.

b) pour l'utilisateur d'un système de reconnaissance, la nécessité de pratiquer un apprentissage avant chaque utilisation du système est une contrainte sérieuse, qu'il convient de réduire au minimum. Par contre s'il est fait une fois pour toutes, l'apprentissage peut être aussi long que l'on veut.

VI.1.3. Aspects spécifiques de la reconnaissance vocale

La reconnaissance vocale soulève d'emblée les problèmes les plus ardus de la reconnaissance des formes, problèmes comparables à ceux que pose la reconnaissance de l'écriture manuscrite. Ces problèmes ont trait à la variabilité du message, aux moyens d'analyse, à la définition même des objets de la reconnaissance et à leur segmentation.

VI.1.3.1. Variabilité du message et adaptation homme-machine

a) Variabilité de la parole

Il nous suffira ici de rappeler la notion de champ de liberté de la parole vue au chapitre IV et illustrée par des expériences de synthèse au chapitre V : déjà pour un même locuteur, le même message peut se présenter dans une infinité de configurations différentes, selon le type de voix adopté (voisée, chuchotée, criée, chantée, etc...), selon l'intonation, le débit et le rythme, selon l'information esthétique en général (y compris la proportion de bruit dans le spectre, le niveau, l'anamorphose instantanée ou intonation formantique, etc), dont nous avons vu que le débit était très supérieur à celui de l'information sémantique recherchée.

D'un locuteur à l'autre on retrouve ces mêmes champs de liberté, rendus plus vastes par l'anamorphose fréquentielle et par les différences de prononciation (accent personnel ou familial; accent régional : peut-on vraiment penser que le son (ã) prononcé par un méridional soit objectivement identique au /ã/ d'un parisien ?)

b) Adaptation de la machine au locuteur

Lorsqu'on donne les performances d'un système de reconnaissance, on indique en général le nombre de locuteurs reconnus. Cette précision n'a qu'une valeur très relative, même quand on spécifie qu'il s'agit de voix d'hommes ou de femmes, car les voix peuvent être très voisines ou très différentes à l'intérieur d'un groupe. Une typologie des voix, fondée

...../

sur l'anamorphose plutôt que sur la hauteur du fondamental, fait cruellement défaut. En l'absence de cette typologie, les résultats devraient être présentés sous forme d'une proportion de voix bien reconnues; parmi un groupe assez vaste (une centaine) comprenant des voix d'hommes, de femmes et d'enfants. Cette proportion serait véritablement un indice de la capacité d'adaptation du système.

c) Dérive de la voix et adaptation du locuteur

Lorsqu'on demande à un locuteur de parler devant un système de reconnaissance on peut observer deux types de phénomènes agissant en sens opposé (J.Y. GRESSER bib. VI-6). La dérive de la voix est une modification lente et progressive de ses caractéristiques : un système bien adapté au début à une voix donnée peut voir ses performances se dégrader, alors même que le locuteur s'efforce de parler de manière reproductible. Par contre, si le locuteur connaît bien le système qu'il utilise (et à plus forte raison si ce dernier fonctionne en temps réel), il arrive à compenser certaines lacunes du système, souvent de manière inconsciente. Ce qui, en soi, n'est pas un défaut : tout locuteur cherche à se faire comprendre de son auditeur, fût-il une machine, et pour cela s'adapte aux conditions qui lui sont imposées. Cependant, ce phénomène peut fausser considérablement l'estimation des performances d'un système de reconnaissance; comme en synthèse, il est nécessaire de recourir à des expérimentateurs "naïfs", donc impartiaux. Ce qui procure quelquefois des surprises bien désagréables.

On peut, à l'inverse, avoir parfaitement conscience de ces phénomènes et en tirer une philosophie : s'il semble trop difficile de concevoir un système s'adaptant à toutes les voix, apprenons aux utilisateurs à se faire comprendre de la machine en mettant à profit leur capacité d'adaptation. Nous pensons cependant qu'il ne faut pas pousser cette idée trop loin, car elle est stérilisante : elle conduit par exemple à créer un nouveau langage approprié aux besoins de la reconnaissance automatique (bib. VI-7) au lieu de chercher à améliorer celle-ci. Or, l'échec de l'espéranto, comme l'évolution des langages de programmation, montrent à l'évidence qu'il faut respecter les contraintes des langages existants. Il faut, au contraire, chercher à exploiter toutes les ressources des langues réelles, ressources présentes dans le vocabulaire, la syntaxe et la sémantique. Nous en revenons ainsi à la reconnaissance conçue comme une simulation de la perception.

VI.1.3.2. Moyens d'analyse

Dans la majorité des cas, l'analyse du message est effectuée par un banc de filtres, par exemple par l'organe analyseur d'un VOCODER. Nous avons vu que cet appareil présentait certains défauts, imputables en partie au détecteur de mélodie. Aussi une utilisation inconsidérée du VOCODER comme organe d'entrée constitue-t-elle un handicap certain, à moins d'augmenter le nombre de canaux, de revoir le recouvrement des bandes passantes, et de ne pas utiliser le signal de mélodie.

Actuellement presque tous les systèmes de reconnaissance utilisent un calculateur; soit en simulation d'appareillages analogiques, soit en application de méthodes purement numériques. Il se pose donc un problème d'échantillonnage : à quels instants faut-il prélever l'information du

...../

signal ? Si ces instants sont trop rapprochés, les volumes de traitement et de mémoire nécessaires risquent d'être trop importants; s'ils sont trop éloignés on risque de perdre une partie importante de l'information (certaines explosions ont une durée de quelques millisecondes).

L'analyse-codage de la parole a trouvé récemment une nouvelle orientation, avec le comptage des passages à zéro du signal (zéro-crossing methods). Nous en verrons un exemple dans la réalisation de VICENS et REDDY (bib. VI-8).

En réalité, les capteurs (ou plutôt "codeurs", l'analyse étant incluse), sont soumis à des impératifs divers, que l'on peut résumer comme suit :

a) Ils doivent préserver l'information utile; leur définition dépend donc des connaissances acquises sur la structure de la parole.

b) Ils doivent éliminer dans la mesure du possible l'information "inutile" ou superflue, pour décharger d'autant le calculateur.

c) Ils doivent être adaptables à des expérimentations variées (divers locuteurs, divers niveaux sonores, divers bruits de fond,etc...) : en bonne gestion, les fonctions d'adaptation doivent être décentralisées au maximum.

d) Ils doivent être réalisables technologiquement, et fonctionner de préférence en temps réel.

En fin de compte, la structure du capteur devrait être déterminée par les opérations à effectuer sur le message, et non l'inverse. On est donc amené à se poser la question : que cherche-t-on à reconnaître ?

VI.1.3.3. Les objets de la reconnaissance

Le cerveau humain possède une curieuse faculté de compréhension à divers niveaux. Nous percevons le message parlé , continu par essence, comme une suite discontinue de mots. Lorsque le contexte nous permet de prévoir les mots, l'opération que nous appelons compréhension d'un mot se réduit à sa comparaison globale avec quelques mots sélectionnés : si le mot est imprévu, sa compréhension nécessite une recherche en mémoire. Mais si le mot est inconnu, nous sommes capables de l'analyser et de le transcrire phonétiquement.

Bien que les systèmes actuels soient incapables de prendre en compte le sens du message, nous utiliserons cette remarque pour distinguer deux approches différentes du problème de reconnaissance automatique.

a) Reconnaissance analytique, ou phonétique

Par ce terme nous entendons la reconnaissance d'éléments phonétiques du message : phonèmes, phonatomes, syllabes, etc.. Le résultat en est une chaîne ininterrompue de symboles phonétiques. Elle ne suppose aucune

...../

limitation du vocabulaire utilisé, mais une prononciation "correcte", c'est-à-dire dans laquelle tous les éléments à reconnaître soient présents.

La segmentation constitue pour ce type de reconnaissance un problème essentiel : le système doit lui-même déterminer les frontières des éléments phonétiques dans le flux continu du message, avant de passer à leur reconnaissance proprement dite (qui comporte une normalisation ou une paramétrisation des éléments).

On peut ultérieurement regrouper cette suite phonétique en mots discrets, moyennant la connaissance du vocabulaire utilisé, chaque mot du répertoire étant stocké en mémoire dans sa notation phonétique; c'est seulement à ce niveau que l'on peut envisager de rectifier d'éventuelles erreurs de prononciation ou de reconnaissance. Mais la suite phonétique peut aussi être codée et transmise telle quelle à un synthétiseur (Icophone, ou synthétiseur commandé par règles). Dans ce cas le débit moyen sur la ligne de transmission est de l'ordre de 50 bits par seconde.

b) Reconnaissance globale, ou "par mots"

On suppose alors le vocabulaire limité à N mots, dont les configurations-types sont mémorisées dans le système sous forme brute ou sous forme paramétrique. Chaque mot à reconnaître est traité dans son ensemble; peu importent les éléments phonétiques qui le composent. On rencontre alors deux problèmes importants :

- La segmentation du message en mots est une opération encore plus difficile que la segmentation en éléments phonétiques. C'est le problème de la "parole continue" (connected speech), que la plupart des chercheurs ont évité jusqu'à présent (à quelques exceptions près) en imposant la prononciation mot par mot (rappelons que le "mot" est entendu ici au sens de groupement phonétique, et peut représenter une expression complète ou une phrase, pourvu qu'elle ait été mémorisée ainsi dans le répertoire).

- Le même "mot", peut présenter d'une occurrence à l'autre des différences considérables de rythme et de durée (anamorphoses temporelles). Toute comparaison avec les références doit donc être précédée d'une normalisation temporelle passablement compliquée, ou, ce qui revient au même, d'une paramétrisation qui ne retienne du message que ses éléments caractéristiques, sous une forme intemporelle.

On voit donc qu'il est impossible, en ce domaine comme en bien d'autres, d'échapper aux problèmes temporels. Malgré cela, la reconnaissance par mots semble devoir aboutir plus rapidement que la reconnaissance phonétique dont l'ambition est plus vaste. Nous allons maintenant examiner rapidement quelques réalisations-types, qui relèvent de l'une ou l'autre de ces démarches.

...../

VI.1.4. Evolution récente de la reconnaissance vocale

Il n'est pas question de passer en revue les très nombreuses réalisations qui ont vu le jour depuis une vingtaine d'années. Nous chercherons seulement à illustrer par quelques exemples l'évolution récente des recherches. Pour une bibliographie plus complète on se reportera par exemple aux références VI-9 et VI-10.

VI.1.4.1. Evolution de la reconnaissance analytique

Dans les premiers temps les recherches sont limitées par la notion de phonème et par les problèmes technologiques (le transistor et l'ordinateur n'ont pas encore suscité de grands bouleversements dans les laboratoires). SMITH (bib. VI-11) présente en 1951 un détecteur de phonèmes. DREYFUS-GRAF (1961, bib. VI-12) met au point son "phonétographe", appareillage analogique composé de 20 filtres passe-bande et de circuits identificateurs de phonèmes. Le phonétographe utilise des "compresseurs sélectifs" qui augmentent l'émergence de certains phonèmes. Le résultat, obtenu en temps réel, est spectaculaire; cependant l'appareil ne fonctionne qu'avec une seule voix, astreinte à une prononciation particulière qui accentue chaque phonème.

FRY et DENES (1958, bib. VI-13) utilisent également un banc de 20 filtres et des masques pour 14 phonèmes. De plus, ils tiennent compte des probabilités de digrammes, qui améliorent le taux de reconnaissance : celui-ci passe de 60 % à 70 % pour un locuteur. Lorsque plusieurs locuteurs participent à l'expérience, le taux de reconnaissance tombe de 70 % à 45 %.

OLSON et BELAR (1961, bib. VI-14) sont parmi les premiers à envisager la reconnaissance d'unités phonétiques autres que les phonèmes. Leurs unités sont des "syllabes phonétiques", que le locuteur doit articuler séparément, ou du moins avec une chute importante du niveau sonore en guise de séparation. Il s'agit donc presque d'une reconnaissance par mots isolés, étant entendu que ces "mots" sont courts, et que leur répertoire est assez limité : d'après DEWEY, 2000 syllabes suffisent à couvrir 98 % des besoins de la langue anglaise. La "Phonetic Typewriter" utilise un banc de 8 filtres répartis selon une loi logaritmique et un dispositif de détection des formants (maxima du spectre, utilisés ensuite en tout ou rien). Chaque syllabe est normalisée en temps : la configuration binaire issue des 8 filtres n'est prise en compte que lorsqu'elle indique une variation spectrale assez importante, ce qui permet d'associer à chaque syllabe une matrice de 5 x 8 positions binaires. On applique ensuite une méthode de reconnaissance par corrélation, avec apprentissage. Le système fonctionne correctement pour un locuteur, mais aurait mérité un plus grand succès. Il semble, vu le faible nombre de filtres, qu'une répartition linéaire eût donné de meilleurs résultats qu'une répartition logarithmique, qui exagère l'importance du premier formant.

M. LAMOTTE et ses collaborateurs utilisent un banc de 35 filtres échelonnés de manière logarithmique, suivi d'une matrice d'apprentissage simulée sur ordinateur (1969, bib. VI-15). Le prétraitement comporte des opérations d'intégration temporelle qui se ramènent à une normalisation de la durée des phonèmes. Le taux de reconnaissance, compté sur les phonèmes, est de 80 % après une trentaine de cycles d'apprentissage sur une même voix. Mais l'optimum n'est pas encore atteint.

...../

J.P. TUBACH (1970, bib. VI-16) cherche à reconnaître un vocabulaire limité (les 81 mots du langage ALGOL 60), prononcé par un seul locuteur, à partir des données fournies par un VOCODER à 20 canaux. Un programme de segmentation décompose les mots en 5 types de segments, selon la présence du voisement, des deux premiers formants, ou d'un bruit de friction. Les segments de type "vocalique" sont caractérisés par l'évolution temporelle des deux premiers formants. La reconnaissance, dans sa phase analytique, nécessite la mémorisation par apprentissage de plusieurs centaines de segments. L'originalité de ce travail réside surtout dans l'utilisation des contraintes syntactico-sémantiques du langage ALGOL 60, qui permettent d'obtenir une bonne reconnaissance des mots à partir d'une reconnaissance rudimentaire (mais performante) des segments.

VI.1.4.2. Evolution de la reconnaissance globale

La réalisation de DAVIS, BIDDULPH et BALASHEK est, à notre connaissance, la première à aborder la reconnaissance de manière globale (1952, bib. VI-17). Les dix chiffres "zéro" à "nine" sont reconnus analogiquement, avec un bon taux de réussite pour une seule voix, à partir de la configuration obtenues dans le plan F_1 F_2. Les fréquences F_1 et F_2 ne sont pas mesurées, mais approchées par le comptage des passages à zéro du signal dans deux bandes (inférieure et supérieure à 900 Hz).

DENES et MATTHEWS (bib. VI-18), pour reconnaître les 10 premiers chiffres, comparent globalement les représentations temps-fréquence, numérisées et normalisées en durée totale. Le taux d'erreur est nul pour un seul locuteur, et s'élève à 6 % pour 5 locuteurs ayant participé à l'apprentissage.

La réalisation de VICENTS (1969, bib. VI-8) est parmi les plus remarquables de ces dernières années. Sa réussite est due en grande partie à un processus de segmentation très élaboré, qui détecte la variation d'ensemble de 6 paramètres déterminés toutes les 10 ms. Ces paramètres sont le nombre des passages à zéro et l'amplitude maxima dans trois bandes de fréquence (150-900 Hz; 900-2200 Hz; 2200-5000 Hz). A chaque segment, caractérisé par 6 paramètres, est affectée une classe phonétique parmi 13 possibles. L'ensemble d'un mot est représenté par une suite d'une dizaine de segments, qui donne lieu à une matrice de paramètres (feature matrix). Chaque mot-référence est mémorisé sous cette forme. La procédure d'identification est extrêmement souple et opère une sélection progressive parmi les mots-références. En cas de mauvaise identification, il est prevu une procédure d'apprentissage corrigé. Le système donne d'excellents résultats (entre 90 et 100 %) pour un locuteur, en trois ou quatre cycles d'apprentissage, même pour un vocabulaire étendu (jusqu'à plus de 500 mots).

Le "Voice Command System" de GLENN et HITCHCOCK est la première réalisation commerciale en reconnaissance vocale (1971, bib. VI-19). C'est un appareil autonome qui reconnaît de manière fiable 24 mots isolés, moyennant 5 cycles d'apprentissage par le même locuteur. L'analyse du message est effectuée par un banc de 16 filtres. Chaque mot est représenté par 8 événements, prélevés aux instants de plus grande variation interne du message. Cette normalisation temporelle, ainsi que les traitements d'apprentissage et de reconnaissance, est confiée à un petit calculateur incorporé.

Pour terminer ce tour d'horizon, nous mentionnerons SAKOE et CHIBA (1971, bib. VI-20), qui abordent le problème de la parole continue. Ils effectuent

...../

une normalisation temporelle implicite par le moyen d'une programmation dynamique : les petites différences de temps entre la configuration à reconnaître et la référence sont compensées à l'instant même de la comparaison, par la recherche de l'événement-référence le plus comparable à l'évènement incident. Ce principe nécessite une consultation permanente des mots-références, qui ne peuvent donc pas être trop nombreux, mais permet en contrepartie de reconnaître des mots dans le flot continu de la parole. Un taux de 99,2 % a été atteint dans la reconnaissance de 500 nombres de deux chiffres (japonais) prononcés par 5 locuteurs masculins, avec pour seules références les 10 nombres de 1 chiffre.

VI.1.4.3. Bilan des recherches actuelles

Il semble que la reconnaissance analytique ait conduit à une grande désillusion, sans doute parce que le problème a été posé de manière trop simpliste. La notion de phonème, en tant qu'invariant acoustique, et la théorie de l'Information, dans la mesure où elle sous-estime les problèmes syntaxiques et sémantiques, nous paraissent être les causes principales de ce demi-échec. Cette voie de recherche, séduisante par son universalité, ne doit pas être abandonnée, mais au contraire explorée avec des hypothèses plus réalistes, tenant compte du caractère dynamique de la parole, et tolérant un certain pourcentage d'erreur qui doit diminuer à mesure que l'on simule un niveau supérieur de la perception.

La reconnaissance globale de mots isolés, mémorisés par apprentissage, a actuellement pris le pas sur la reconnaissance analytique, car elle semble plus rapidement opérationnelle dans l'environnement des calculateurs. Elle ne doit cependant pas faire illusion à son tour : la normalisation temporelle est aussi difficile que la segmentation en éléments ; l'apprentissage, même interactif, la séparation des mots et la limitation du vocabulaire sont d'importantes contraintes imposées à l'utilisateur.

Personne n'a, jusqu'à présent, résolu de manière satisfaisante le problème d'adaptation à un locuteur quelconque. Il faut d'ailleurs remarquer que, très généralement, les résultats de reconnaissance portent sur des mots prononcés de manière comparable, en rythme, niveau et intonation; nous ne connaissons pas de système qui travaille indifféremment sur la voix voisée, chuchotée, criée ou chantée, même pour un seul locuteur.

Nous allons maintenant présenter quelques expérimentations qui ne prétendent pas, pour l'instant, atteindre de très bonnes performances, mais qui peuvent illustrer de nouvelles voies d'approche des problèmes de reconnaissance.

VI.2. CODAGE NUMERIQUE ET ANALYSE SPECTRALE DE LA PAROLE

Les expérimentations que nous avions prévues nécessitaient l'introduction de parole en machine. Ne disposant pas au CIRCE d'un convertisseur analogique-numérique, nous avons dû effectuer cette opération ailleurs, c'est-à-dire au GRI (Groupe de Recherches Ionosphériques du CNRS, à ST-MAUR), grâce à l'amabilité de M. MOURETON, Mme PARIS et M. CROZES. Une analyse-conversion utilisant un VOCODER

...../

déjà existant eût sans doute été plus facile à réaliser; mais nous préférions déterminer nous-mêmes les principaux paramètres de l'analyse (nombre et répartition des voies, cadence d'échantillonnage, etc..), de façon à obtenir finalement une représentation temps-fréquence satisfaisante, au regard de notre expérience en spectrographie acoustique et en synthèse de la parole.

Cette partie de l'étude s'est faite en plusieurs étapes : définition et enregistrement du corpus de parole, codage analogique-numérique, analyse par transformée rapide de FOURIER (FFT), pondérations et lissages, et enfin édition sur imprimante. La majeure partie de la programmation a été assurée par M. MLOUKA (bib. VI-21).

VI.2.1. Définition et enregistrement du corpus de parole

Les voix de 8 locuteurs, masculins et féminins, ont été enregistrées au laboratoire d'Acoustique. On n'a pas cherché à réaliser des enregistrements de très haute qualité, de façon à rester dans les conditions usuelles de la transmission acoustique : la dynamique est de l'ordre de 30 dB, et il existe une très légère réverbération. Aucune contrainte de prononciation ou de débit n'a été imposée aux locuteurs, si ce n'est la recommandation de parler à haute et intelligible voix. Néanmoins, dans l'ensemble, le débit de parole est un peu plus faible que le débit observable dans la conversation courante, sauf pour trois des locuteurs, habitués à parler devant un microphone.

La durée totale de l'enregistrement est de l'ordre de 45 minutes. Les différentes parties peuvent être classées comme suit :

VI.2.1.1. Signaux de référence

- Sinus 1000 Hz, niveau nominal 0 dB
- Sinus 1000 Hz, niveau décroissant depuis 0 dB jusqu'à - 40 dB par bonds de 5 dB
- Bruit blanc, niveau nominal 0 dB
- Signal carré 100 Hz, niveau nominal 0 dB
- Sinus, 0 dB, fréquence croissante 30 - 10 000 Hz.

VI.2.1.2. Grands fichiers (1 voix d'homme, 1 voix de femme)

a) Phrase : "Les petits oiseaux chantent, c'est le printemps", en voix voisée avec et sans intonation, et en voix chuchotée.

b) Ensemble de séquences du type "APABATADAKAGA...", "ACHAJASAZA..." etc.., comprenant tous les phonatomes utiles de la langue française.

c) Phrases constituées à partir d'un nombre réduit de phonatomes représentatifs :

...../

- Le petit chat gris a pris la fuite
- Marie ramasse une cerise
- Je suis un hippophage
- Arthur a un ami bizarre
- Tintin a lu le fascicule sur la Birmanie
- Babar marche sur le bureau
- As-tu vu ce fameux lapin ?

d) Tous les mots, isolés, constituant ces phrases.

VI.2.1.3. Petits fichiers (3 voix d'homme, 3 voix de femme)

Le contenu est identique à 2 c) et 2 d); de plus on a demandé à ces 6 locuteurs de prononcer un dictionnaire réduit comportant tous les phonatomes utilisés dans les phfrases et les mots.

L'enregistrement est effectué sur une bande magnétique standard (1/4 de pouce, 19 cm/s, pleine piste) avec un enregistreur NAGRA. Chaque séquence est précédée d'un claquement bref destiné au repérage ultérieur.

VI.2.2. Codage analogique-numérique

Le GRI dispose d'un convertisseur analogique-numérique commandé par un calculateur PDP5. Le signal analogique, lu à partir d'un enregistrement à modulation de fréquence, est transformé en une suite d'échantillons numériques et mémorisé sur bande. Ce processus, bien que très simple dans son principe, s'est avéré en pratique très fastidieux et générateur d'erreurs, les appareillages et les codes utilisés étant trop divers et mal adaptés les uns aux autres.

VI.2.2.1. Préparation de l'enregistrement analogique

Après montage, il a fallu transposer l'enregistrement en modulation de fréquence, sur bande 7 pistes 1/2 pouce (enregistreur AMPEX FR 1800 L) et ajouter en synchronisme sur deux autres pistes deux autres signaux. Le premier donnait la cadence d'échantillonnage (signal sinusoïdal 10 000 Hz, de façon à définir une fréquence maxima de 5000 Hz), et le second était un signal en créneau permettant de n'effectuer la conversion que sur les séquences utiles, en éliminant les tops de repérage et les silences entre phrases. Malgré les difficultés présentées par cette seconde opération, entièrement effectuée à la main, son efficacité est hors de doute, puisqu'elle a permis de réduire l'enregistrement de 45 mn à 30 mn utiles, tout en délimitant à 0,5 seconde près le début et la fin de chaque séquence.

A ce niveau nous avons omis d'effectuer un filtrage passe-bas, qui eût éliminé du signal les fréquences supérieures à 5000 Hz. Cette inobservation du théorème des échantillonnages entraîne un intéressant retournement du spectre, que nous commenterons plus loin.

...../

VI.2.2.2. Résultat du codage

Chacun des 10 000 échantillons par seconde est codé sur 2 caractères de la bande numérique 7 pistes, soit 12 bits et deux bits de contrôle. Le codage est binaire pur positif, de 0 à 4095. Les échantillons sont groupés par blocs de 1400 valeurs séparés par des "gaps". Les séquences sont séparées les unes des autres par des blocs mis à zéro, et un numéro d'ordre est écrit au début de chaque séquence. L'ensemble occupe 3 bandes 7 pistes 556 bpi à noyau compatible IBM, et chaque bande se termine par un "End of File".

A titre de contrôle, le début de la première bande numérique a été relu et décodé par le PDP5, et a donné lieu à un tracé sur enregistreur graphique CALCOMP (à partir de 1 point sur 10). La fig. VI-12 reproduit un des oscillogrammes obtenus. Nous avons pu vérifier ainsi que le signal n'avait pas subi d'altération visible, mises à part certaines interruptions aléatoires dues à une mauvaise lecture analogique du signal en créneau.

VI.2.2.3. Décodage sur IBM 360 du CIRCE

Cet ordinateur fonctionnant avec des mots de 4 octets, M. MLOUKA a dû écrire un programme spécial de décodage en PL1. Nous avons été amenés à imprimer l'ensemble des données acoustiques, à raison de une valeur sur cent et à repérer manuellement les numéros de blocs du début et de la fin de chaque séquence. Cette opération a permis par ailleurs d'éliminer les instants de silence résiduels, et de rectifier les erreurs dans la numérotation des séquences, erreurs qui provenaient des interruptions signalées plus haut.

Nous avons constaté, durant cette manipulation, que le volume de données dont nous disposions était énorme (environ 2.10^7 échantillons de longueur 12 bits), et par conséquent peu maniable en machine, alors que la version écrite en phonétique des données occupait quelques pages. Cette remarque illustre concrètement la faible proportion d'information sémantique présente dans le message parlé.

VI.2.3. Analyse spectrale

Il s'agissait d'obtenir, de manière entièrement numérique, un diagramme temps-fréquence comparable à celui que donne le Sonagraphe en bande large ($\Delta F = 300$ Hz). Nous avons utilisé pour cela un programme FFT, aimablement communiqué par M. DELOUIS (CIRCE).

VI.2.3.1. Détermination de la fenêtre temporelle

Comme nous l'avons vu au chapitre II, les caractéristiques de l'analyse sont essentiellement déterminées par la durée et la forme du prélèvement temporel. Après de nombreux essais, nous nous sommes arrêtés aux caractéristiques suivantes :

...../

a) Durée du prélèvement : 12,8 ms, soit 128 échantillons, qui conduisent à un spectre de puissance défini par 64 points en fréquence, échelonnés entre 0 et 5 000 Hz avec un pas de 78,2 Hz.

b) Forme du prélèvement : pour éviter des discontinuités artificielles d'un spectre à l'autre, il est nécessaire de réduire l'amplitude des échantillons au début et à la fin du prélèvement. Ceci est effectué en multipliant chaque échantillon par une fonction de pondération (fenêtre de Hanning) de la forme

$$H(k) = \frac{1}{2} \left(1 - \cos \frac{2 \pi k}{128} \right)$$

où k est le numéro de l'échantillon, allant de 0 à 127.

c) Recouvrement des prélèvements : toujours pour atténuer les discontinuités temporelles, nous avons effectué les prélèvements toutes les 10 ms, soit avec un recouvrement de 2,8 ms (28 échantillons) d'un prélèvement sur le suivant.

VI.2.3.2. Analyse par FFT

La grandeur qui nous intéresse est l'amplitude spectrale. C'est le module $|A_r|$ du coefficient complexe A_r fourni par la FFT :

$$A_r = \sum_{k=0}^{127} X_k \exp(-2 \pi j r k / 128)$$

X_k est le $k^{ième}$ échantillon temporel : $k = 0,1 \ldots 127$

A_r est le $r^{ième}$ coefficient fréquentiel : $r = 0,1 \ldots 127$

La fonction à analyser étant réelle, $|A_r|$ est une fonction paire, et il suffit de considérer les valeurs de r comprises entre 64 et 127. La résolution obtenue était trop grande pour que les formants apparaissent nettement; il a fallu effectuer un lissage du spectre, en remplaçant chaque valeur par la moyenne de trois valeurs consécutives.

Après ces traitements, la largeur apparente d'analyse, difficile à mesurer ou à calculer précisément, à cause de la quantification, peut être estimée à

$\Delta F \simeq 250$ Hz à - 3 dB.

VI.2.3.3. Retournement du spectre

La figure VI-10 illustre le retournement du spectre qui provient de l'omission du filtrage passe-bas à 5 kHz. Le signal référence de fréquence

...../

croissante, que nous croyions quasi-sinusoïdal, était en fait affecté d'une distorsion harmonique (présence d'harmonique 3) assez importante, imputable au générateur utilisé. Lorsque le fondamental dépasse 1,67 kHz, l'harmonique 3 dépasse 5 kHz et apparaît en miroir dans le spectre. Malgré ce défaut, nous avons conservé nos données telles quelles, pensant que l'énergie de la parole au-delà de 5 kHz était négligeable. Ce phénomène ne pouvait guère se manifester que pour /s/ et /f/, et nous avons estimé qu'il y avait peut être là un moyen - bien involontaire - de détecter avec une analyse limitée à 5 kHz des sons de fréquence supérieure.

VI.2.4. Pondération fréquentielle

C'est un fait bien connu que l'énergie moyenne de la voix décroît de 6 à 10 dB par octave, au-delà d'une fréquence située entre 500 et 1000 Hz, ceci avec des variations assez considérables d'un locuteur à l'autre. Il est donc nécessaire d'appliquer une pondération renforçant les valeurs spectrales situées au-dessus de 1000 Hz. Il faut, de plus, tenir compte du maximum de sensibilité de l'oreille entre 1000 et 3000 Hz. Cette zone correspond sensiblement au domaine d'évolution du 2e formant, qui est le plus significatif. Après divers essais (bib. VI-22) nous avons mis au point une pondération fréquentielle satisfaisante en première approximation pour les 8 locuteurs : les valeurs spectrales situées au-dessous de 500 Hz sont pratiquement inchangées; vers 2500 Hz elles sont multipliées par 10 et vers 5000 Hz par 4 (figure VI-11).

En réalité, chaque locuteur possède un timbre particulier, pour un type de voix et un niveau sonore donnés. C'est dire que la pondération fréquentielle devrait être individualisée, autour de la valeur moyenne définie ci-dessus. Il semble, d'après la doctrine de E. LEIPP, concernant l'oreille moyenne (cf chapitre III), que la nature ait résolu ce problème par le moyen de l'adaptation ossiculaire : la courbe de réponse de la chaîne ossiculaire présente des résonances variables avec le degré de tension des muscles du marteau et de l'étrier, résonances que l'auditeur règle inconsciemment de façon à faire émerger au mieux les formes acoustiques significatives.

Un tel processus, requérant une rétroaction "intelligente", semble difficile à simuler actuellement; mais une possibilité d'adaptation sélective sera évoquée au paragraphe VI.3.1.

VI.2.5. Représentation imprimée

Le sonagramme étant un document visuel, il fallait, pour mettre au point l'analyse et la segmentation, élaborer une représentation numérique équivalente, qui permît de conserver le caractère global des formes acoustiques. Divers essais de noircissement par superposition de caractères ont été effectués, à la suite desquels un simple codage en 5 nuances a été adopté : les valeurs spectrales étant ramenées à l'intervalle 0-10 on a imprimé

- pour 0 un blanc
- pour 1 l'apostrophe '
- pour 2 le chiffre 1
- pour 3 le chiffre 2
- au-dessus de 3 : l'astérisque *

..../

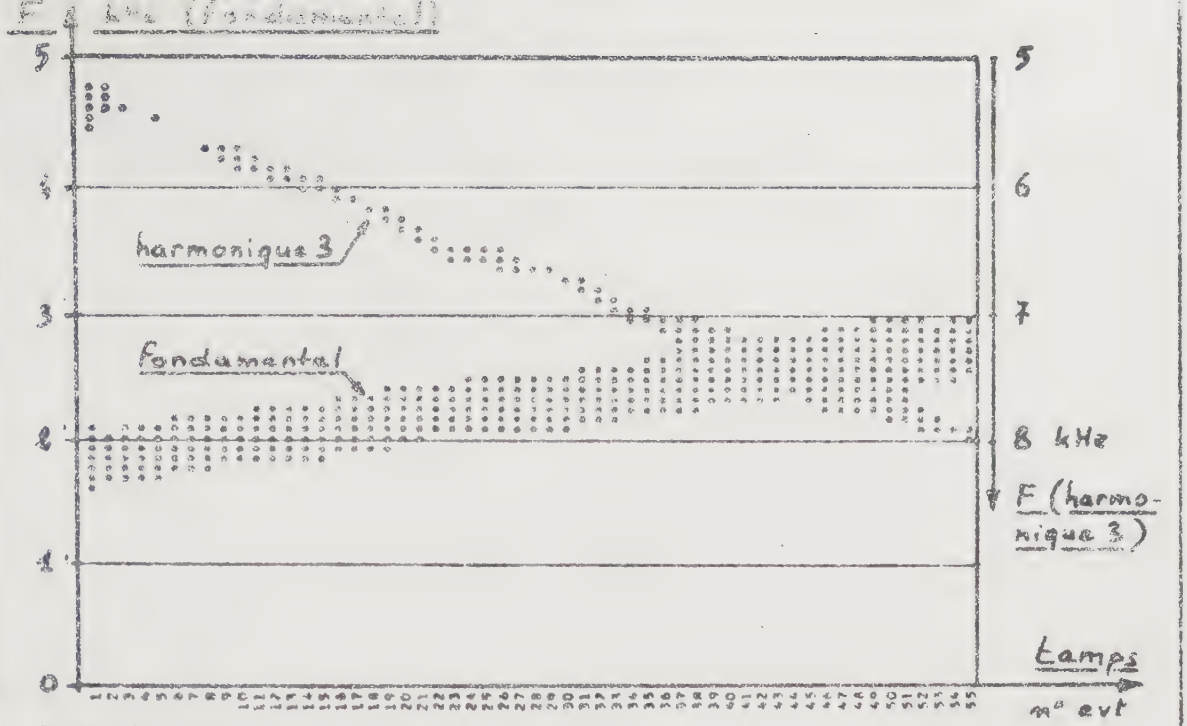


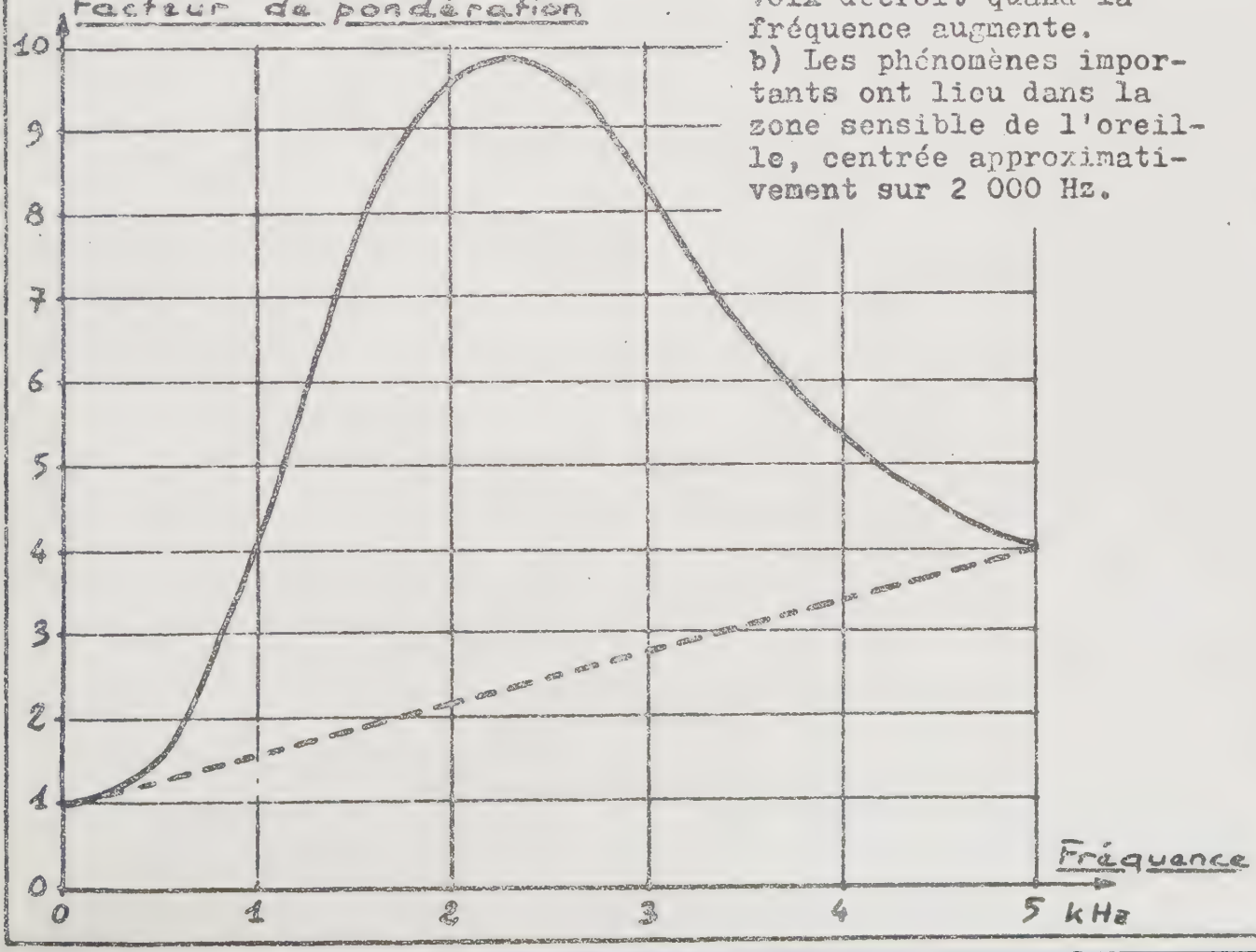
Fig VI-10

Repliement du spectre autour de la fréquence de NYQUIST (ici 5 000 Hz). Les fréquences supérieures apparaissent en miroir : le fondamental croît en fréquence, alors que l'harmonique 3 semble décroître.

Fig VI-11

Pondération fréquentielle utilisée dans les essais de reconnaissance. Cette correction, appliquée à l'amplitude spectrale issue de l'analyse, résulte de deux sortes de considérations :

a) L'énergie moyenne de la voix décroît quand la fréquence augmente.
b) Les phénomènes importants ont lieu dans la zone sensible de l'oreille, centrée approximativement sur 2 000 Hz.



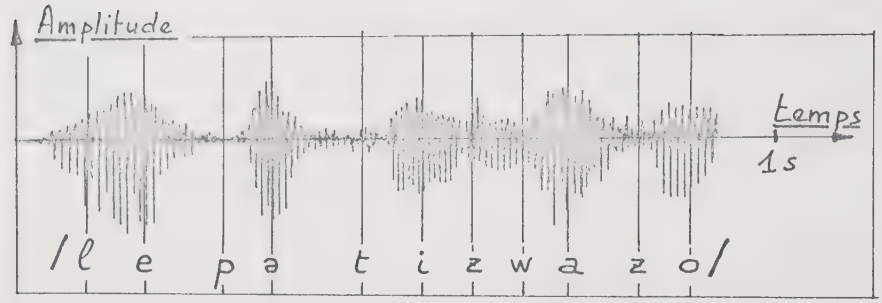


Fig VI-12

Oscillogramme de la séquence "Les petits oiseaux", après conversion analogique-numérique et relecture (locuteur masculin : JSL).

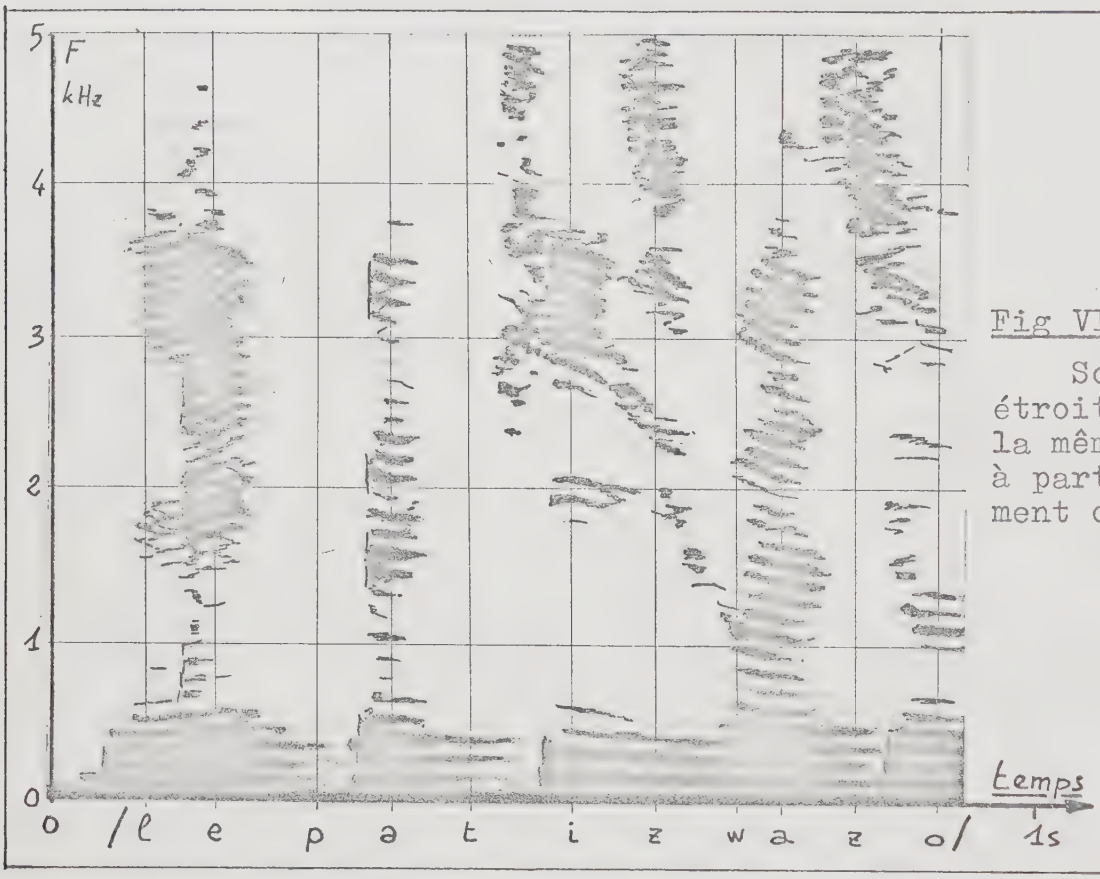


Fig VI-13

Sonagramme (en bande étroite : ΔF =45 Hz) de la même séquence, obtenu à partir de l'enregistrement original.

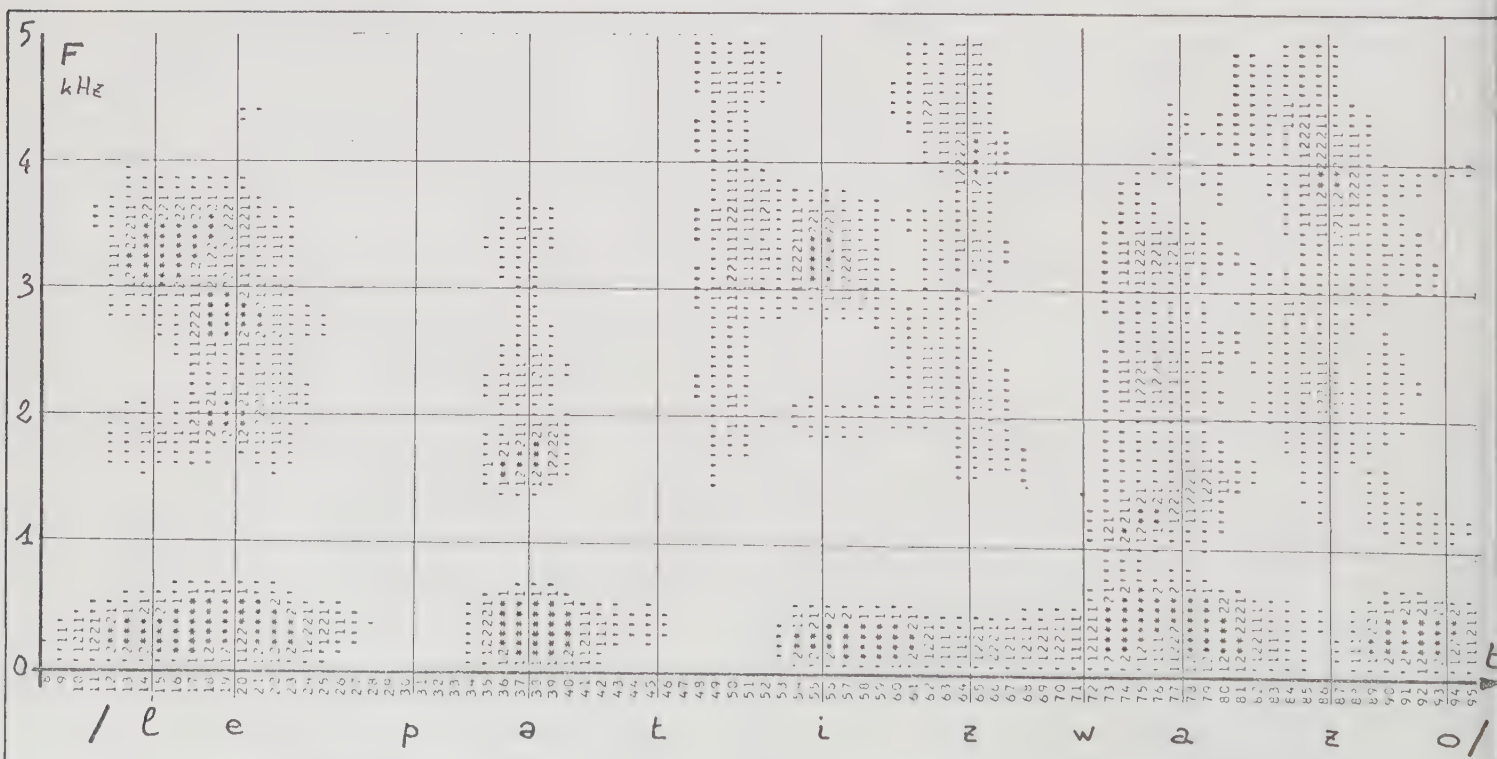


Fig VI-14 Sonagramme numérique de la même séquence, obtenu par transformation rapide de FOURIER. Après pondération fréquentielle et lissage, la largeur apparente d'analyse est de l'ordre de 250 Hz.

La figure VI-14 représente le sonagramme numérique du début de la phrase " Les petits oiseaux chantent, c'est le printemps ", prononcée par un locuteur masculin. Outre les traitements décrits plus haut, on a ici effectué un lissage temporel sur trois spectres successifs, et imprimé seulement une valeur sur deux en fréquence. Les formes acoustiques sont parfaitement nettes et comparables à celles du sonagramme correspondant (fig. VI-13).

Du point de vue de la programmation, le volume de données à traiter et à mémoriser a posé de gros problèmes. Il a fallu utiliser un "disk-pack" et plusieurs heures de calcul sur IBM 360 75 pour analyser l'ensemble du corpus. Mais actuellement chacune des 585 séquences du corpus acoustique est disponible sous forme de sonagrammes numériques, moyennant quelques secondes de calcul en unité centrale.

Nous avons, grâce à l'obligeance de M. JAMATI et des Ingénieurs de l'Institut BLAISE PASCAL, pu utiliser un lecteur optique élaboré, pour numériser des sonagrammes. Cette voie aurait mérité un approfondissement, car elle présentait certains avantages (utilisation du Sonagraphe). Nous l'avons néanmoins laissée de côté, car le codage et l'analyse FFT étaient à ce moment pratiquement opérationnels.

VI.3. NORMALISATION DYNAMIQUE ET FREQUENTIELLE DE LA PAROLE

Avant d'aborder la reconnaissance proprement dite, il nous faut encore examiner deux types de traitements à appliquer à la parole pour réduire son champ de liberté.

VI.3.1. Normalisation dynamique

Le niveau sonore de la parole est une grandeur extrêmement variable : un même locuteur peut parler plus ou moins fort, s'approcher du microphone, tourner la tête, etc...; le phénomène se produit a fortiori d'un locuteur à l'autre, et les variations de niveau sonore atteignent rapidement 6, 12, voire 20 dB, ce qui correspond à des rapports de 2, 4 ou 10 en amplitude spectrale. Ces variations, à peine remarquées par l'oreille, sont particulièrement gênantes en reconnaissance automatique, car les configurations à comparer doivent être normalisées en niveau.

Le problème est moins aigu si les configurations sont paramétrisées, p.ex. si l'on s'intéresse à l'évolution des fréquences de formants; mais nous avons vu plus haut qu'une tentative d'extraction automatique des paramètres semblait prématurée, en l'état actuel des connaissances. Si l'on tient à conserver le caractère global de la configuration temps-fréquence, on peut effectuer la normalisation dynamique soit sur l'ensemble de la configuration à étudier (ce qui ne peut se faire qu'en différé), soit instantanément sur chaque événement. Nous avons choisi la seconde méthode pour la reconnaissance analytique, pour pouvoir la transposer éventuellement dans le domaine analogique.

La régulation dynamique doit à notre sens avoir un double but :

- elle doit maintenir à peu près constant le niveau sonore moyen, estimé sur une période d'au moins 50 à 100 ms en parole courante

...../

- elle doit accentuer certaines modifications spectrales significatives du message, comme les bruits d'impact rencontrés dans les plosives. Il est nécessaire pour cela de tenir compte du niveau sonore du message, non seulement à l'instant présent, mais aussi dans le passé immédiat. La correction tend par exemple à augmenter le niveau si le message est faible, mais ceci d'autant plus qu'il est faible depuis plus longtemps.

La fig. VI-15 donne le schéma analogique d'une telle rétroaction dans laquelle le signal f (t) est régulé à l'aide d'une ligne à retard; les signaux f (t - τ) issus de la ligne sont pondérés selon une fonction h (τ), avant d'être additionnés pour donner le signal de régulation r (t). Le gain g(t) du régulateur varie en fonction inverse de r (t). On effectue ainsi l'opération

$$f'(t) = \frac{f(t)}{\int_0^{\infty} h(\tau)\, f(t-\tau)\, d\tau}$$

On notera la parenté formelle de ces opérations avec celles que nous avons attribuées au cerveau (chapitre III); l'analyse temps-fréquence et l'autocorrélation requièrent les mêmes composantes : retards, sommations, inhibitions.

Ce processus a été simulé de la manière suivante : le signal à réguler est en fait la somme à l'instant i des amplitudes spectrales de la configuration X_{ij} :

$$S_i = \sum_j X_{ij} \qquad \begin{matrix} (\ i = 1,2, \ldots\ N\ , \text{ indice de temps} \\ ) \\ (\ ) \ j = 1,2, \ldots\ 64\ , \text{ indice de fréquence} \end{matrix}$$

On définit une fonction de pondération H_k portant sur K événements passés (fig. VI-16), dont la forme dépend de l'importance relative que l'on veut donner au présent et au passé immédiat; ce qui permet d'obtenir la fonction de régulation R_i

$$R_i = \sum_{k=0}^{k=K} H_k \cdot S_{(i-k)}$$

Après quoi chaque valeur spectrale est divisée par un coefficient variant en fonction de R_i

$$X'_{ij} = \frac{X_{ij}}{A.\ R_i + B}$$

A et B sont des constantes permettant d'ajuster le fonctionnement du régulateur.

L'effet de ce processus est visible sur la figure VI-17 qui représente le même début de séquence, sans et avec régulation. On notera en particulier l'évolution des courbes :

...../

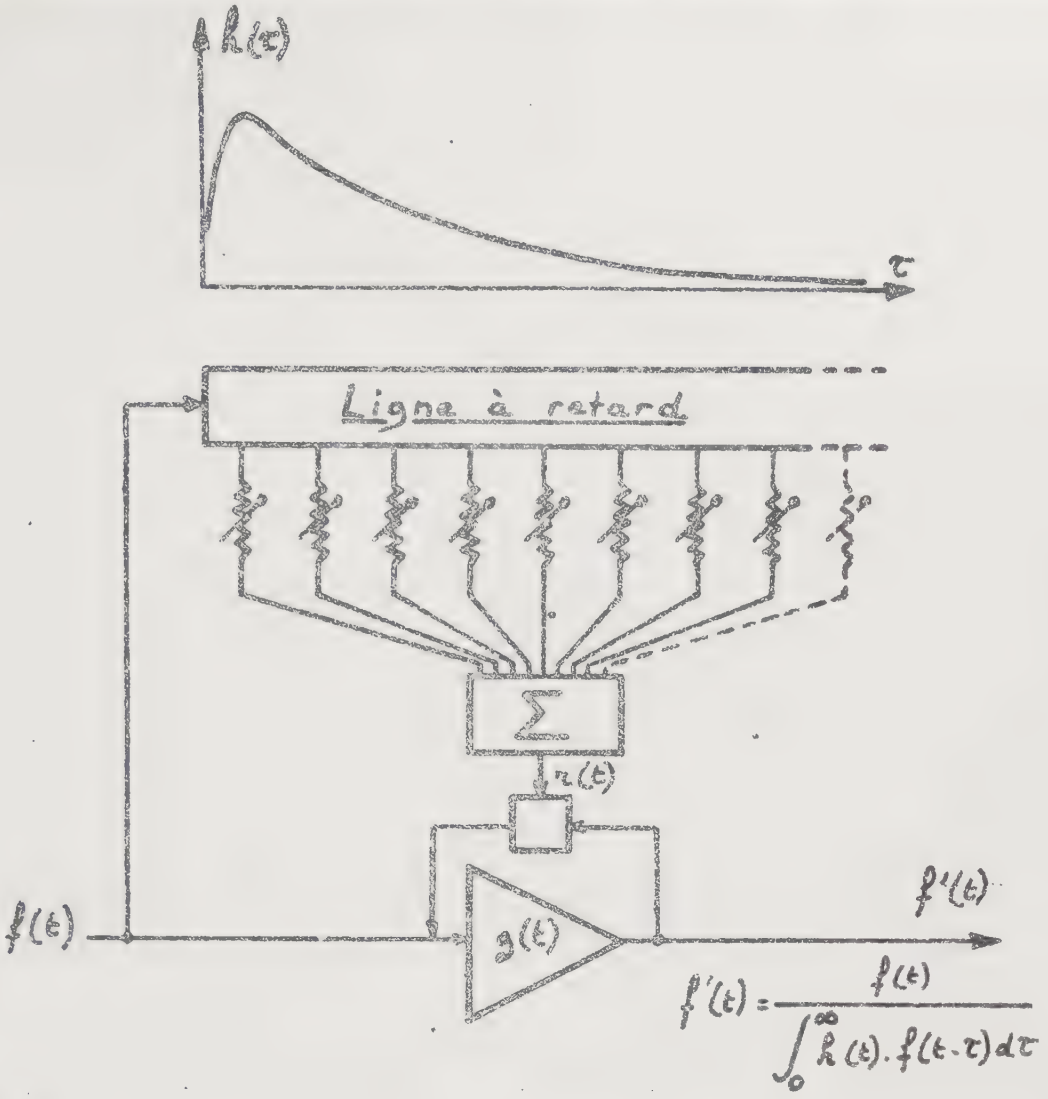


Fig VI-15

Schéma analogique de la régulation dynamique retardée. La courbe $h(\tau)$ représente l'influence relative des événements passés ; elle est matérialisée par des coefficients ajustables. Le signal de régulation agit sur la ligne de rétroaction d'un amplificateur.

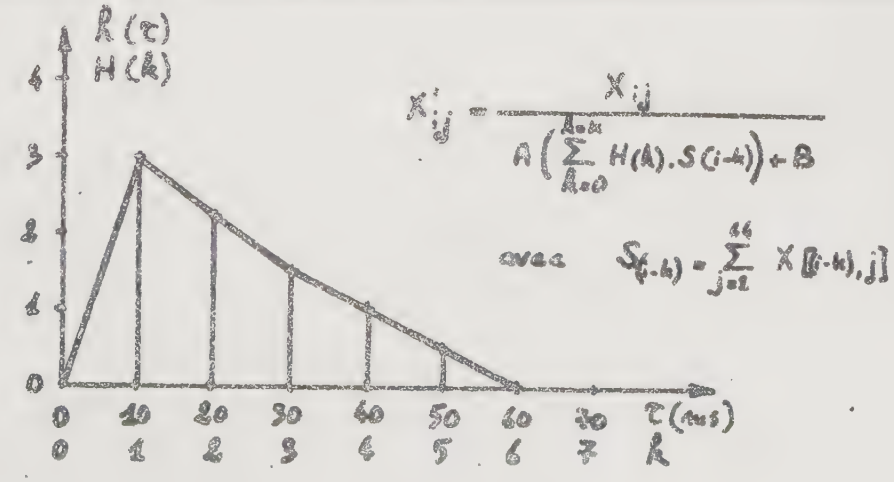


Fig VI-16

Valeurs H(k) adoptées dans la simulation, le signal à réguler étant la somme par événement S_i. En fait, la correction est appliquée à chaque valeur spectrale X(i,j).

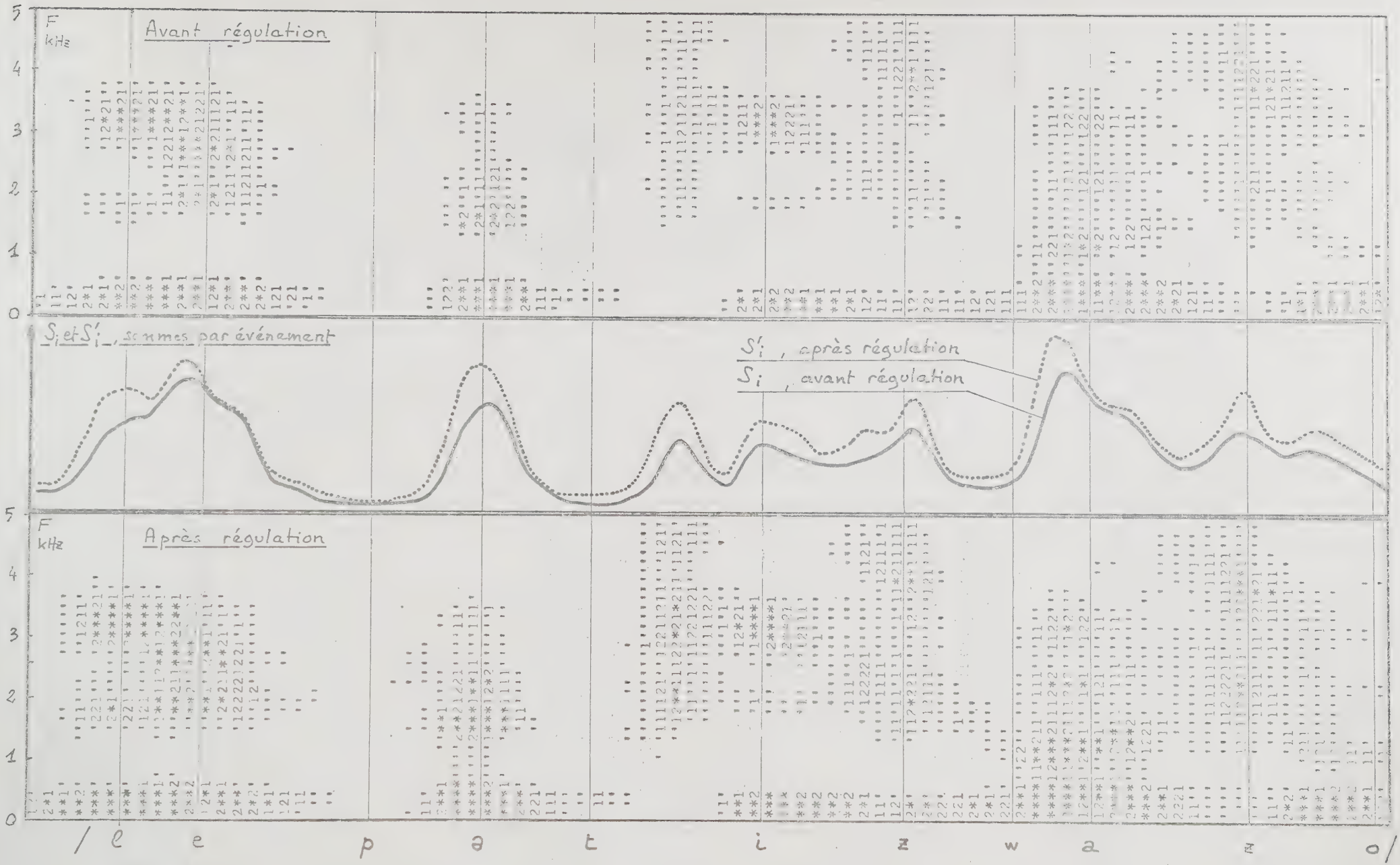


Fig VI-17

Effet de la régulation dynamique sur la séquence "Les petits oiseaux" (locuteur JSL). La régulation augmente le niveau moyen de la séquence, mais la correction est d'autant plus forte qu'elle s'applique après un silence ou un son faible (voir par exemple /l/, /ə/, /ti/ et /wa/). Réciproquement, le niveau diminue après quelques 30 ms de son fort (sur /e/ et /a/, pas exemple).

$$S_i = \sum_j X_{ij} \qquad \text{et} \qquad S'_i = \sum_j X'_{ij}$$

Le but visé est bien atteint : dans l'ensemble les valeurs X_{ij} sont légèrement renforcées, mais ce renforcement est inégalement réparti dans le temps, intervenant surtout lors de l'attaque des sons.

Nous avons pu obtenir de cette manière des séquences de niveau sensiblement homogène pour les huit locuteurs. Néanmoins, ce processus doit faire l'objet d'une étude plus approfondie, guidée par les résultats objectifs de la reconnaissance. En particulier il serait intéressant de l'appliquer non pas à l'ensemble du message, mais à plusieurs bandes de fréquence, ce qui permettrait d'envisager une régulation automatique du timbre de la parole. On effectuerait ainsi une simulation fonctionnelle plus complète de l'oreille moyenne et de sa commande nerveuse.

VI.3.2. Normalisation fréquentielle

Nous avons déjà détaillé le problème des anamorphoses fréquentielles existant entre locuteurs. Nous nous proposons de l'illustrer ici par la détermination numérique du coefficient d'anamorphose optimum.

VI.3.2.1. Détermination du coefficient d'anamorphose optimum

La séquence R (t,F), provenant d'un locuteur pris comme référence, est comparée à la séquence X (t,F), de même contenu phonétique, mais prononcée par un autre locuteur. La comparaison donne lieu à un taux de ressemblance ou proximité τ_1 . On forme ensuite la séquence anamorphosée X (t,kF) dans laquelle k , coefficient d'anamorphose fréquentielle régulière, peut prendre des valeurs comprises entre 0,5 et 2 (le sous-programme AFREQ, de A. CALINET, effectue cette opération). A chaque valeur de k on associe une valeur τ_k résultant de la comparaison avec la référence, et l'on cherche la valeur de k pour laquelle τ_k est maximum.

Il est nécessaire que les séquences à comparer soient normalisées en temps : nous avons fabriqué 8 séquences normalisées en sélectionnant manuellement dans la séquence " Marie ramasse une cerise " 20 événements choisis dans la partie stable de chaque phonème (voyelle ou consonne) ceci pour chaque locuteur. Deux de ces séquences font l'objet de la fig. VI-18, qui comporte également la courbe $\tau = f(k)$ de l'une par rapport à l'autre. La comparaison est faite par le sous-programme COMPA (annexe II).

Le programme ADAPT (M. MLOUKA) effectue cette opération 64 fois, en prenant tour à tour chaque séquence comme référence, et en la comparant à elle-même et aux 7 autres. Le fait remarquable est que chacune des courbes obtenues passe par un maximum unique, dont l'abscisse correspond au coefficient optimum d'anamorphose. La figure VI-19 montre 8 de ces courbes établies à partir d'une même référence.

Le tableau donné fig. VI-20 met en parallèle les résultats précédents

...../

(coefficients optima) avec ceux fournis par une version antérieure du programme ADAPT, qui travaillait sur de tout autres données acoustiques : les séquences normalisées étaient constituées de 30 évènements choisis sur les voyelles / a œ i o y / et sur leurs transitions. Le pas du paramètre k était de 0,05, et les coefficients optima étaient déterminés sans interpolation, donc à $\pm$ 0,025 près. La concordance entre les résultats peut être considérée comme bonne, sauf pour le dernier locuteur, pour lequel la valeur $k_{opt} = 1,45$ est manifestement fausse : cette erreur provient d'un trop grand écart dynamique entre la séquence donnée et la référence. Notons encore que les coefficients optima trouvés pour les locuteurs MC et EL par rapport à JSL(de l'ordre de 0,90) correspondent parfaitement aux valeurs déterminées manuellement, lors de l'étude anamorphotique des voyelles.

La forme de la courbe $\tau = f(k)$, ainsi que l'ordonnée du maximum (taux optimum de coïncidence), fournissent une indication sur la parenté acoustique de la voix étudiée avec la voix référence. Il serait intéressant de déterminer pour chaque voix les facteurs (anamorphose irrégulière, modification du timbre, etc...) qui permettent d'augmenter le taux optimum. La voix normalisée mise au point par synthèse pourrait être utilisée comme référence. Cette méthode devrait permettre une étude objective de certains aspects esthétiques (plus exactement diagnostiques) de la voix.

VI.3.2.2. Utilisation de l'anamorphose en reconnaissance.

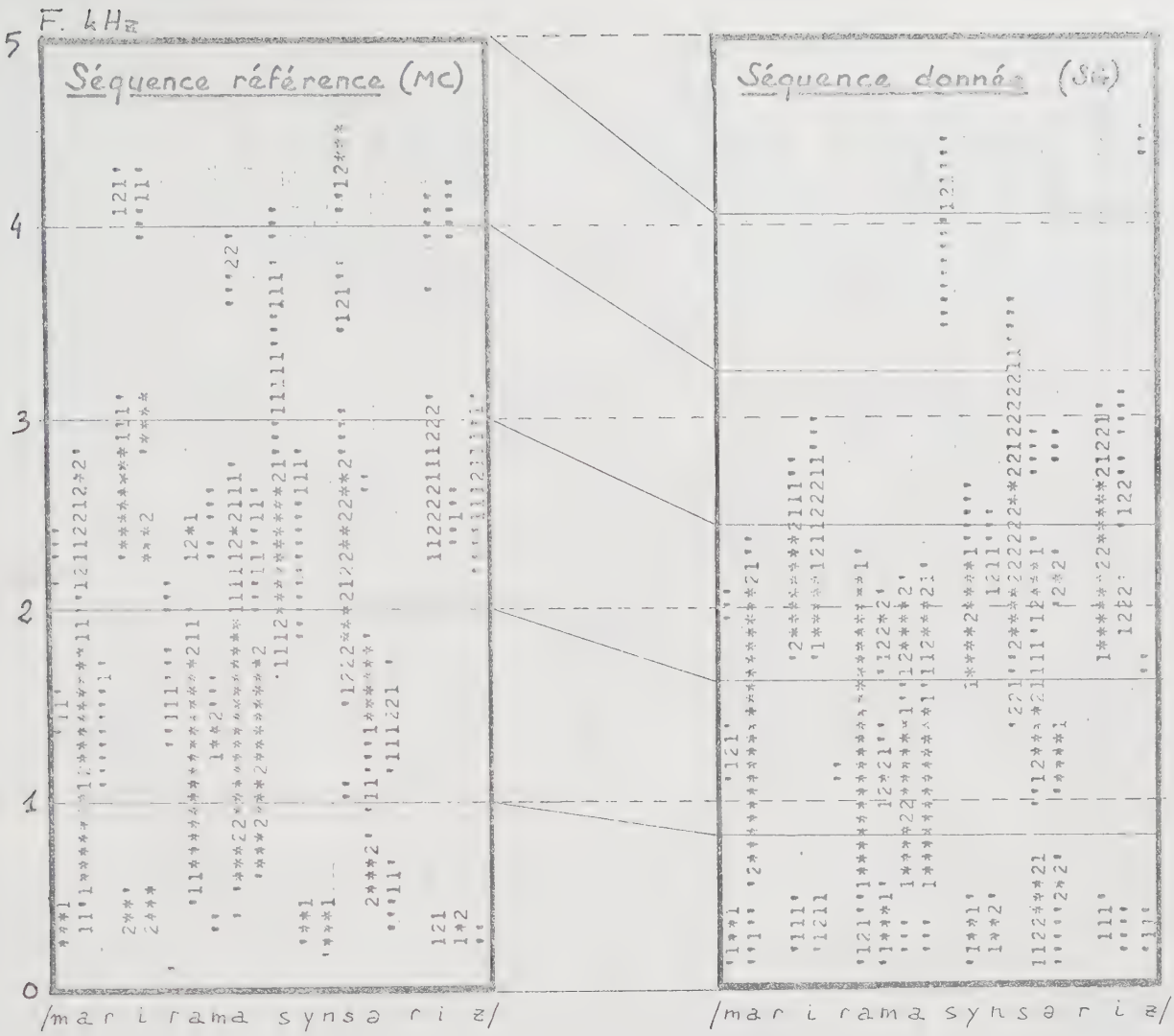
L'anamorphose fréquentielle préoccupe depuis longtemps les chercheurs du Laboratoire d'Acoustique de l'Université PARIS VI (bib. III-13, IV-12, VI-24), et nous avons largement insisté dans les chapitres précédents sur cette notion parfaitement originale. C'est sans doute la clé de l'adaptation au locuteur, en reconnaissance automatique, problème jusqu'à présent non résolu. En effet, connaissant le coefficient optimum d'anamorphose d'une voix particulière par rapport à la voix référence mémorisée dans le système de reconnaissance, il suffit d'appliquer cette anamorphose pour normaliser en fréquence le squelette sémantique de la parole à reconnaître. Cette opération peut être effectuée dès le départ au moyen de l'analyseur-anamorphoseur étudié au Laboratoire de Mécanique Physique de l'Université PARIS VI (J. SAPALY, G.M. MARCOUX, bib. VI-23) : les fréquences centrales de tous les filtres d'analyse sont multipliées par le même coefficient déterminé soit par une expérimentation préalable, soit par un processus interactif avec le système de reconnaissance, si celui-ci fonctionne en temps réel. En simulation, il nous suffit d'utiliser le sous-programme AFREQ lors de la lecture sur disque des données de l'analyse, avec le coefficient d'anamorphose approprié.

VI.3.2.3. Une application de l'anamorphose irrégulière.

Nous avons vu que l'oreille semblait présenter dans la zone sensible sa plus grande finesse en fréquence et intensité. Nous avons vu également que les messages acoustiques les plus efficaces, du point de vue de la perception, étaient en général centrés dans cette zone, et nous avons montré que tel était le cas de la parole.

Pour simuler ce phénomène, nous avons effectué sur nos données de parole une anamorphose irrégulière dont la courbe caractéristique est

...../



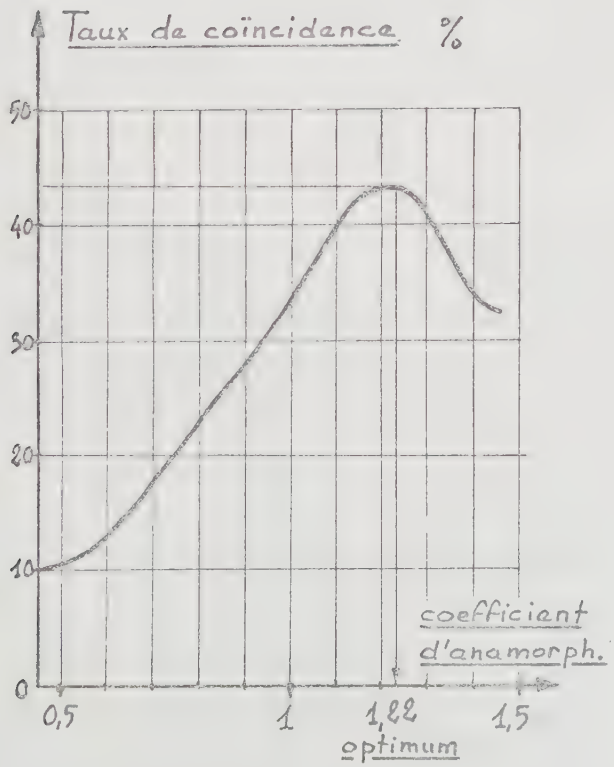


Fig VI-18

Recherche du coefficient optimum d'anamorphose régulière, pour la voix du locuteur SG (masculin) par rapport à celle du locuteur MC (féminin). Les deux séquences ("Marie ramasse une cerise"), normalisées en temps manuellement, sont représentées avec la même échelle fréquentielle. La seconde est anamorphosée en fréquence selon un coefficient variant de 0,5 à 1,5. Le taux de coïncidence passe par un maximum très net, pour le coefficient d'anamorphose optimum de 1,22. Les correspondances fréquentielles ainsi définies sont portées sur les sonagrammes.

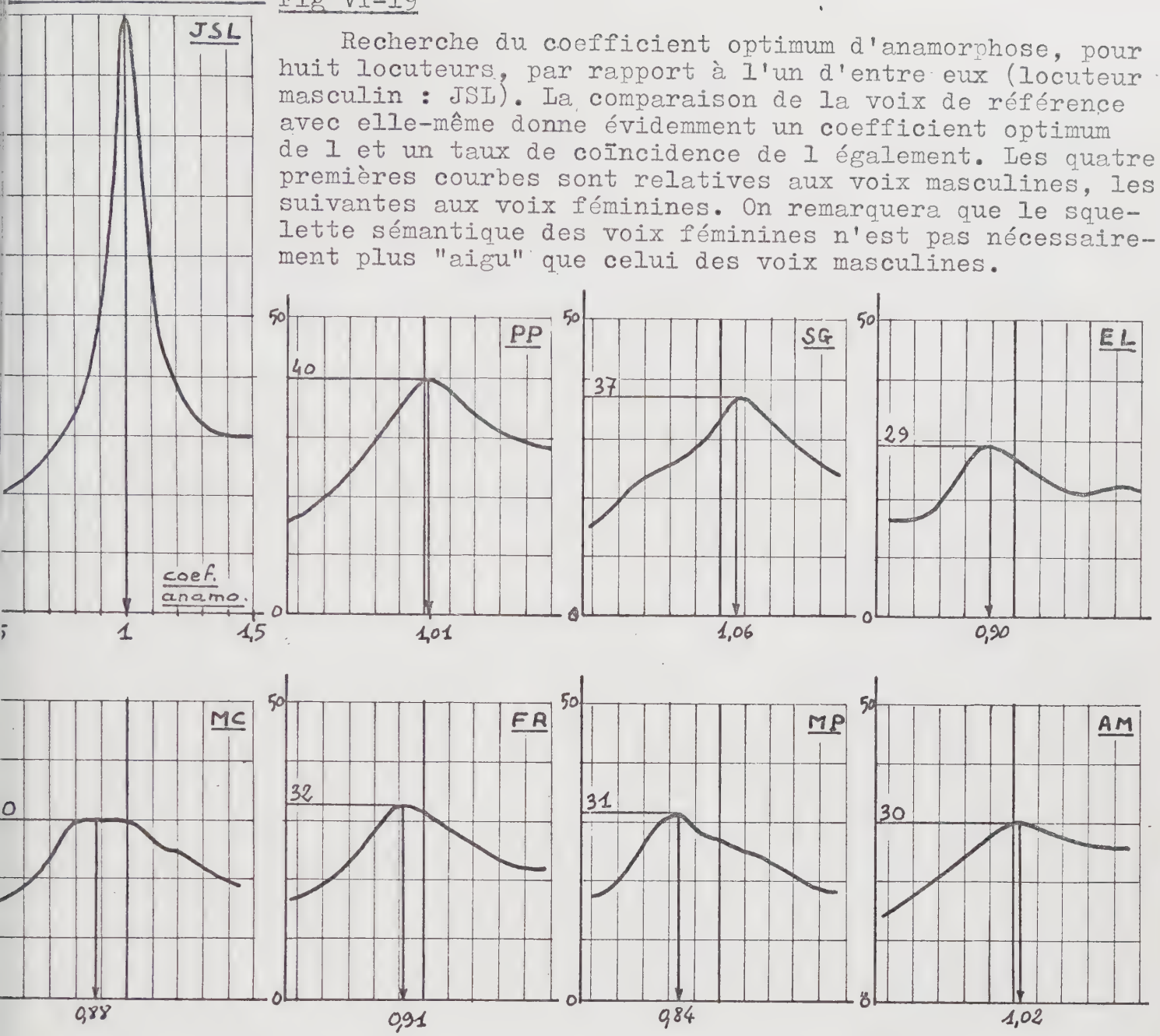


Fig VI-19

Recherche du coefficient optimum d'anamorphose, pour huit locuteurs, par rapport à l'un d'entre eux (locuteur masculin : JSL). La comparaison de la voix de référence avec elle-même donne évidemment un coefficient optimum de 1 et un taux de coïncidence de 1 également. Les quatre premières courbes sont relatives aux voix masculines, les suivantes aux voix féminines. On remarquera que le squelette sémantique des voix féminines n'est pas nécessairement plus "aigu" que celui des voix masculines.

Locuteur	première détermination (sur voyelles uniquement)	seconde détermination (sur une séquence complète)
JSL (réf.)	1,00	1,00
PP	0,95	1,01
SG	1,10	1,06
EL	1,45	0,90
MC	0,85	0,88
FR	0,90	0,91
MP	/	0,84
AM	1,05	1,02

Fig VI-20

Résultats obtenus lors de deux déterminations différentes du coefficient optimum d'anamorphose. La première valeur obtenue pour EL est erronée.

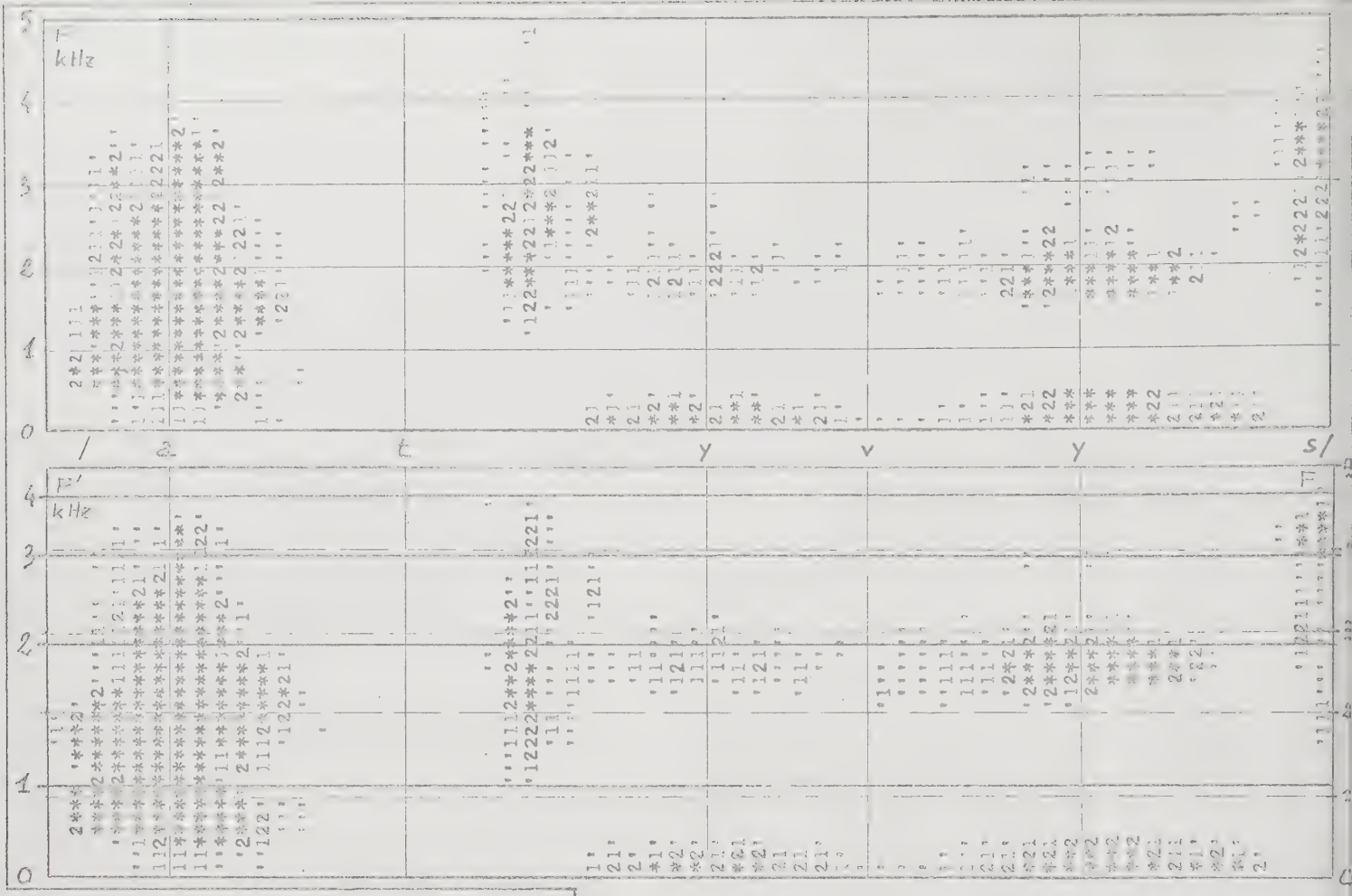
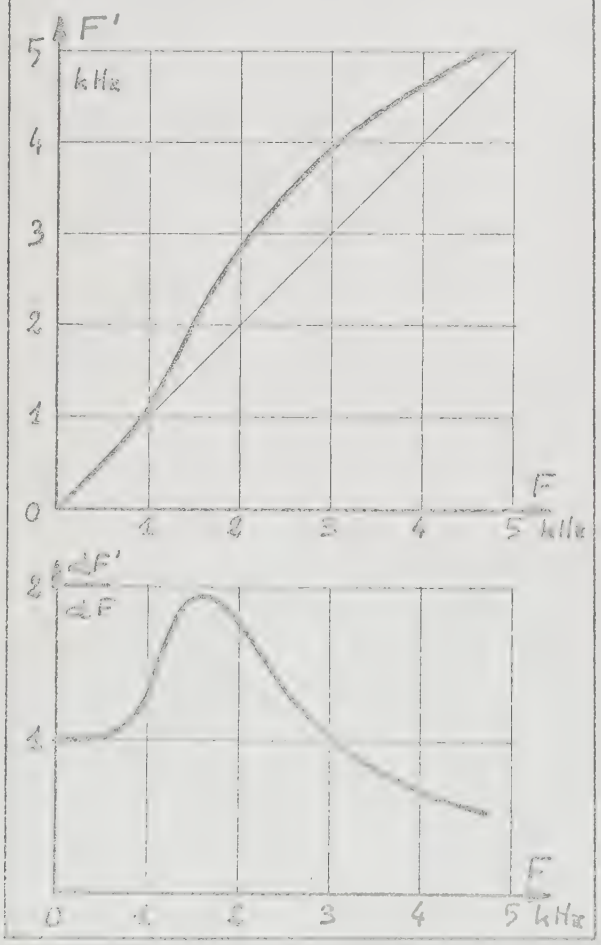


Fig VI-22 Anamorphose fréquentielle irrégulière, appliquée au début de la séquence "As-tu vu ce fameux lapin" (locuteur JSL). La transformation est telle que la zone sensible (1,0-2,0 kHz) soit considérablement dilatée par rapport à la réalité.

Fig VI-21 Courbes caractérisant l'anamorphose irrégulière de la figure VI-22. La dérivée représente la densité spectrale des seuils différentiels.



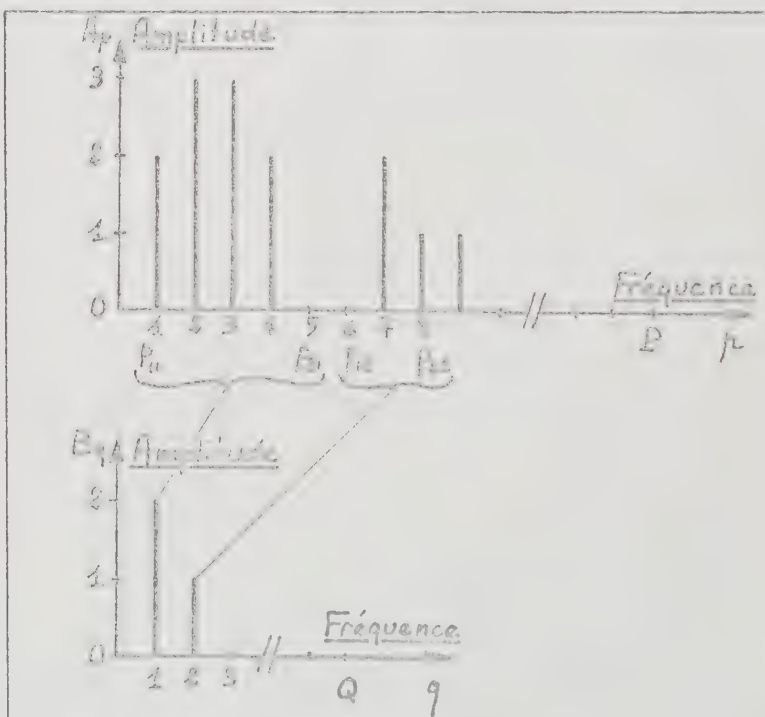


Fig VI-23 Principe du sous-programme ADELE qui, en regroupant les voies d'analyse, effectue simultanément l'anamorphose et le lissage fréquentiel.

$$F' = g(F) \qquad \text{(fig. VI-21)}$$

La nouvelle échelle de fréquence n'est ni linéaire ni logarithmique; sa variation, représentée par la dérivée $\frac{dF'}{dF}$, a la même allure que celle des seuils différentiels de l'oreille en fréquence et intensité.

Le résultat acoustique est visible sur la fig. VI-22, qui représente un sonagramme numérique de parole, avant et après cette transformation : la largeur attribuée aux fréquences 1000-3000 Hz s'est accrue aux dépends des extrémités de la bande (cela revient à examiner à la loupe la partie centrale du sonagramme). Il en résulte une distinction beaucoup plus aisée des fréquences comprises dans cette zone. L'évolution du second formant, déterminante pour le squelette sémantique, est ainsi plus facile à suivre. Nous appliquerons cette transformation lors de la reconnaissance par mots.

Le sous-programme ADELE, qui effectue l'anamorphose irrégulière, permet de passer de $P = 64$ voies d'analyse à $Q < P$ ($Q = 32$ dans l'exemple de la fig. VI-22). Soient Ap les valeurs spectrales avant transformation ($p = 1, 2 \ldots P$), et B_q après ($q = 1, 2, \ldots Q$). Il suffit d'indiquer en tête du programme les indices $p(q)$ des valeurs Ap que l'on veut regrouper dans une seule valeur Bq. Par exemple (fig. VI-23), la valeur B_1 est la moyenne des valeurs A_1, A_2, A_3, A_4 et A_5; la valeur B_2 est la moyenne de A_6, A_7 et A_8, etc.. Si $p_1(q)$ et $p_2(q)$ sont les indices de la première et de la dernière valeur A_p à condenser dans la valeur B_q, on effectue l'opération

$$B_q = \frac{\sum_{p_1(q)}^{p_2(q)} Ap}{p_2(q) - p_1(q) + 1} \qquad \left(q = 1, 2 \ldots Q\right)$$

La souplesse de ce procédé est très grande : l'expérimentateur peut jouer sur les coefficients $p_1(q)$ et $p_2(q)$ pour simuler la répartition et le recouvrement des "pseudo-filtres" donnant les valeurs B_q.

Il serait intéressant, par la suite, de généraliser cette approche en considérant chacune des voies B_q comme une combinaison linéaire des voies A_p :

$$B_q = \sum_{p=1}^{p=P} \alpha_{pq} A_p \qquad q = 1, 2 \ldots Q$$

On simulerait ainsi un banc de filtres "quelconques", défini par la matrice $\{\alpha_{pq}\}$, dont les propriétés spectrales et temporelles seraient très différentes de celles des filtres-résonateurs, et qui pourrait fournir un modèle de la cochlée. Ce modèle contribuerait peut-être à résoudre le problème de l'analyse cochléaire, qui semble se faire suivant plusieurs constantes de temps. Il serait en tous cas important d'étudier la matrice $\{\alpha_{pq}\}$ au vu des résultats objectifs de la reconnaissance.

...../

VI.4. RECONNAISSANCE ANALYTIQUE

La reconnaissance automatique des éléments de parole que sont les phonatomes implique en premier lieu leur normalisation temporelle. Les processus d'apprentissage et de reconnaissance adoptés ensuite sont extrêmement simples dans leur principe. L'organigramme d'ensemble du programme RECO est indiqué fig.VI-24.

VI.4.1. Normalisation temporelle des phonatomes

VI.4.1.1. Principe de la segmentation

La segmentation se déduit très naturellement de la courbe de stabilité du squelette sémantique (mais la segmentation-normalisation que nous pratiquons, comme les notions de squelette sémantique, de phonatome et d'anamorphose qui en sont à l'origine, est à notre connaissance, parfaitement inédite).

Le phonatome est la partie du squelette sémantique comprise entre deux maxima de stabilité. Nous appelons taux de stabilité S(t) le taux de coïncidence (voir annexe II) entre deux événements $E\left(t - \frac{\theta}{2}\right)$ et $E\left(t + \frac{\theta}{2}\right)$ séparés par une durée θ ; S (t) est compris entre 0 et 1 (pratiquement entre 0 et 100 pour éviter les décimales). Nous avons vu au chapitre V que l'allure plus ou moins accidentée de la courbe S (t) caractérisait la netteté d'élocution.

Les maxima relatifs de la courbe ne correspondent pas exactement aux phonèmes de la phonétique classique, mais aux parties stables de ceux-ci. On trouve ainsi une segmentation au milieu de la phase de compression des plosives, ou de la phase d'écoulement aérien des nasales, ainsi que dans toute partie quasi-stationnaire du message (voyelle, transition très lente, silence, certains bruits d'explosion prolongée, etc...). C'est pourquoi nous préférons parler d'événements-frontières (sous-entendu : de phonatomes) plutôt que de phonèmes.

Les minima de la courbe S(t) présentent également un intérêt considérable : ils correspondent aux instants de plus grande instabilité du message. Nous les appellerons événements-pivots et nous les placerons au centre des phonatomes normalisés.

VI.4.1.2. Mise au point de la segmentation-normalisation

La première difficulté consiste à déterminer la durée ou "distance d'ordre" θ . Trop courte (quelques millisecondes), elle ne permet de détecter que les variations très rapides du squelette sémantique, c'est-à-dire pratiquement les explosions. Trop longue (une centaine de milli-secondes), elle atteint la durée moyenne des phonatomes et s'applique donc à un autre niveau phonétique. La chose se complique si l'on veut appliquer la même distance d'ordre à des messages prononcés plus ou moins rapidement : nous avons vu que le débit moyen de parole peut varier dans une très large mesure (de l'ordre de 10 entre extrêmes). Il semble, ici encore, que l'oreille soit

..../

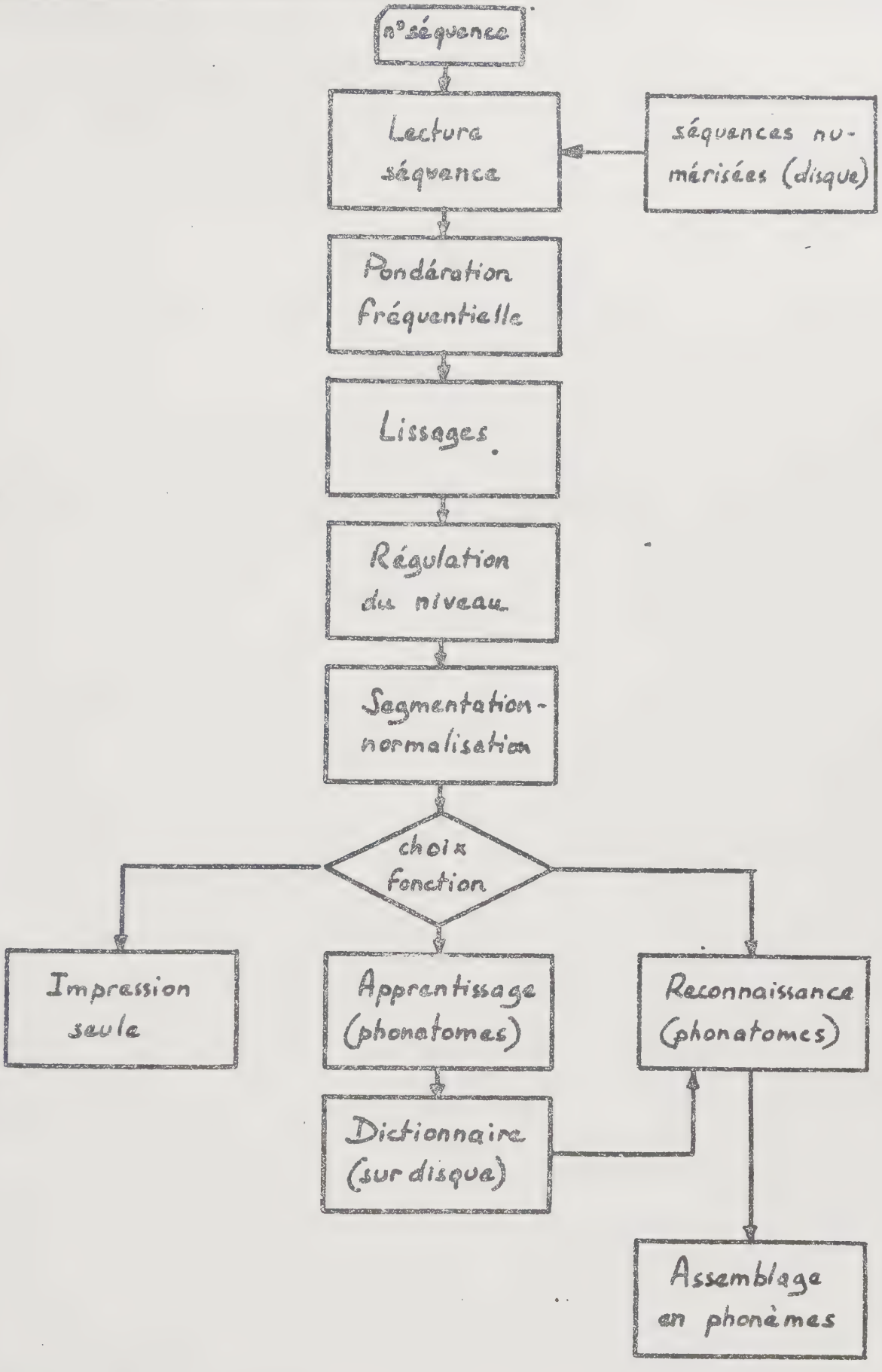


Fig VI-24

Organigramme d'ensemble du programme RECO. Ce programme assure la lecture des séquences stockées sur disque, le prétraitement (y compris la segmentation et la normalisation des phonatomes), et trois fonctions au choix : impression seule du sonagramme et des phonatomes normalisés, apprentissage, ou reconnaissance.

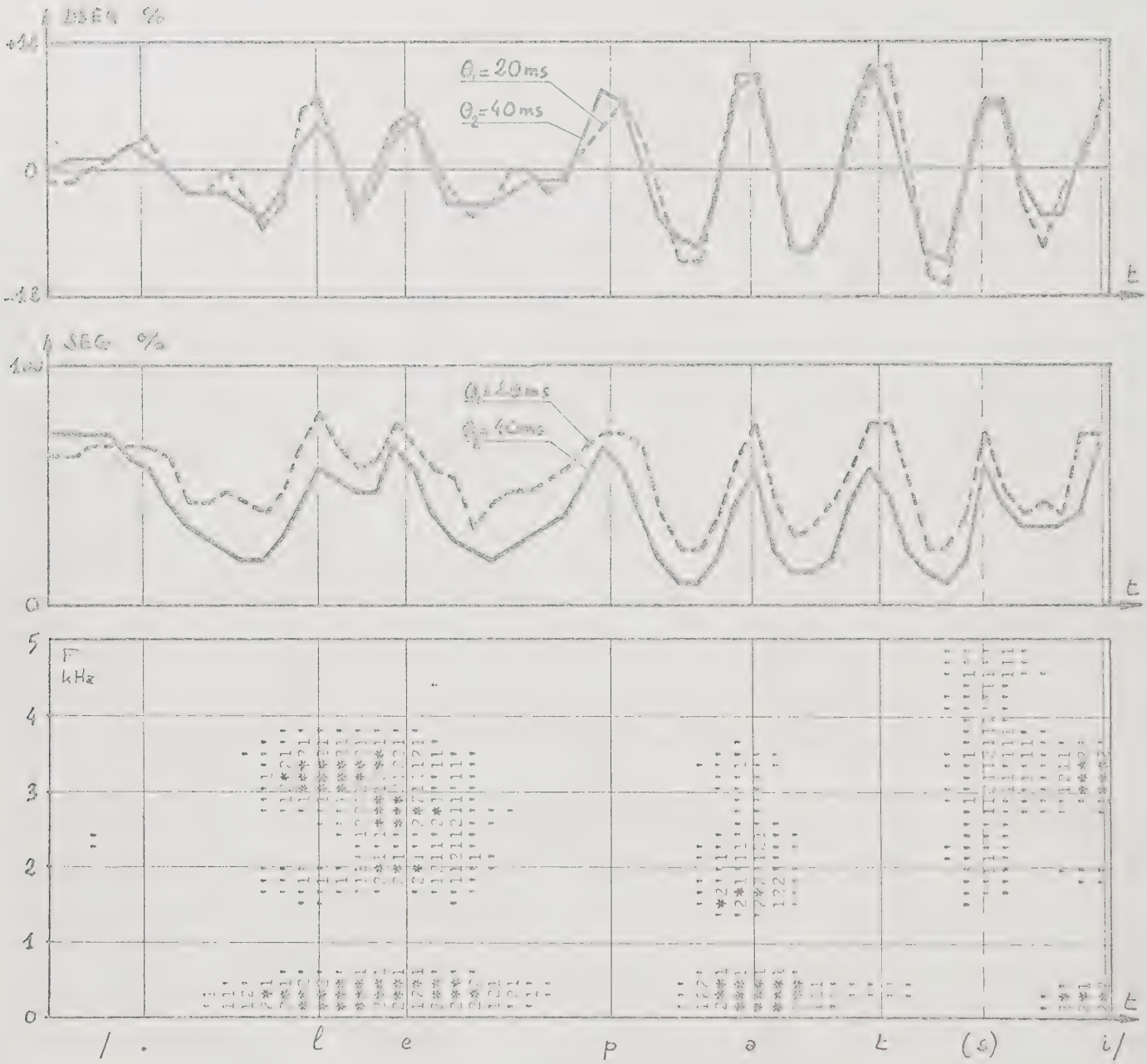


Fig VI-25

Courbe de stabilité de la séquence "Les petits" (locuteur JSL), selon deux valeurs du délai θ. La courbe de stabilité S(t) (ici SEG) présente sensiblement la même forme dans les deux cas ; mais sa valeur moyenne est plus élevée lorsque θ décroît. La courbe DSEG est déduite de SEG par un filtrage digital passe-bande ; ses extrema sont faciles à déterminer. L'explosion du /t/ est considérée ici comme un état quasi-stationnaire, et donne lieu à une segmentation excédentaire qui n'est pas très gênante pour la suite de la reconnaissance. Le spectre à cet instant est d'ailleurs voisin de celui du /s/.

susceptible d'estimer la stabilité des messages selon une gamme de distances d'ordre allant de 1 ms à plusieurs centaines de ms. Cette remarque corrobore celle que nous avons faite à propos de la largeur ΔF d'analyse de l'oreille, qui semble prendre plusieurs valeurs différentes au même instant. Il s'agit d'ailleurs formellement du même phénomène, lié à l'auto-corrélation dans les lignes à retard neuroniques.

La fig. VI-25 représente les courbes S(t) obtenues sur le même message, avec les durées $\theta_1 = 20$ ms et $\theta_2 = 40$ ms. Ces deux courbes sont comparables quant à leurs extrema relatifs; néanmoins la première est plus proche de 1, et les minima sont marqués plus nettements; les caractères opposés s'observent sur la seconde.

Nous avons finalement adopté une durée $\theta = 30$ ms (fig. VI-26) convenant à peu près pour tous les locuteurs du corpus. Cette durée tient compte évidemment du traitement effectué sur la courbe S(t) pour détecter la position des extrema, traitement apparenté à une double dérivation (recherche de la courbure maxima, avec son signe, dans une fenêtre de l'ordre de 100 ms qui représente la durée moyenne d'un phonatome). Il s'agit plus exactement d'un filtrage digital passe-bande, résumé par les expressions suivante

$$SIG(i) = \frac{1}{3} \sum_{i-1}^{i+1} SEG(i)$$

$$DSEG(i) = SIG(i) - \frac{1}{5} \sum_{i-2}^{i+2} SIG(i)$$

SEG (i) est la suite des valeurs de S(t), calculées toutes les 10 ms,

DSEG (i) est sa transformée, et SIG (i) une fonction intermédiaire.

Nous avons surtout cherché à mettre en évidence les segmentations "difficiles", celles que l'on observe sur les liquides et les nasales, pour lesquelles le niveau sonore et les fréquences formantiques varient peu.

Le processus de normalisation, indiqué sur la fig. VI-26, consiste à prélever dans le message cinq événements significatifs : les deux frontières, le pivot et deux événements intermédiaires. Ces derniers sont sélectionnés d'après les passages à zéro de la pseudo-dérivée seconde de S(t). On pourrait sans doute se passer des événements intermédiaires pour la reconnaissance; mais ils contribuent à donner aux phonatomes normalisés des proportions comparables à celles des phonatomes de l'Icophone, que nous connaissons bien, ce qui permet de détecter visuellement d'éventuelles erreurs.

La fig. VI-26 montre une segmentation sur l'explosion du /t/, erreur assez fréquente lorsque le /t/ est prononcée de manière accentuée. Ceci n'est pas très gênant, dans la mesure où la partie stable de l'explosion du /t/ ressemble à un /s/. Le groupe /lepəti/ est alors assimilé à /lepətsi/. En supposant correcte la reconnaissance de /ət/, /ts/ et /si/,

...../

il sera facile d'éliminer ultérieurement le /s/ supplémentaire, par des considérations de durée.

Tel que nous l'avons utilisé, le programme fournit des segmentations excédentaires sur les parties stables des voyelles et des fricatives, qui sont dues à la détection de petites variations de la dérivée S"(t), alors que la courbe S(t) elle-même reste en position haute. Ces erreurs sont minimes, car elles conduisent seulement à augmenter le volume du travail de reconnaissance : au lieu d'avoir à reconnaître /babar/, par exemple, on peut avoir /babaaar/. Ce n'est pas nécessairement un inconvénient, en particulier si l'on pense transmettre la parole à un Icophone sous forme de suite phonétique comportant le rythme du message original. On pourrait néanmoins éliminer une partie de ces erreurs, en ne validant une segmentation que si S(t) se trouve au-dessous d'un certain seuil, par exemple 70 %. Nous ne l'avons pas fait, craignant d'introduire ainsi des fautes plus graves (non-détection d'une segmentation sur une nasale ou une liquide).

VI.4.1.3. Résultats

L'efficacité de la segmentation a été estimée de la manière suivante. Nous avons repéré sur les sonagrammes numériques des 7 phrases d'un seul locuteur les positions "idéales" des événements-frontières, ce qui nous a fourni 145 repères, ceci sans savoir a priori quels événements avaient été sélectionnés par le programme. Nous avons ensuite comparé les positions déterminées par le programme à celles que nous avons déterminées à la main, et nous avons classé les erreurs en trois catégories :

a) Fautes "graves" : non-détection d'une frontière. Nous n'en avons trouvé que 3, soit un taux d'environ 2 %. L'une concernait une suite de deux liquides (/rl/), et les deux autres étaient constituées par la non détection de la partie stable de la voyelle /y/ dans /syr/. Le taux de reconnaissance des phonatomes est donc à priori limité à 98 %.

b) Fautes "rattrapables" : présence d'une segmentation excédentaire, dans une partie transitoire. Nous en avons cité plus haut un exemple (détection d'une frontière sur l'explosion du /t/). Ces erreurs, pour être graves, ne sont pas irrémédiables. Nous en avons compté 16, soit un taux de 11 %.

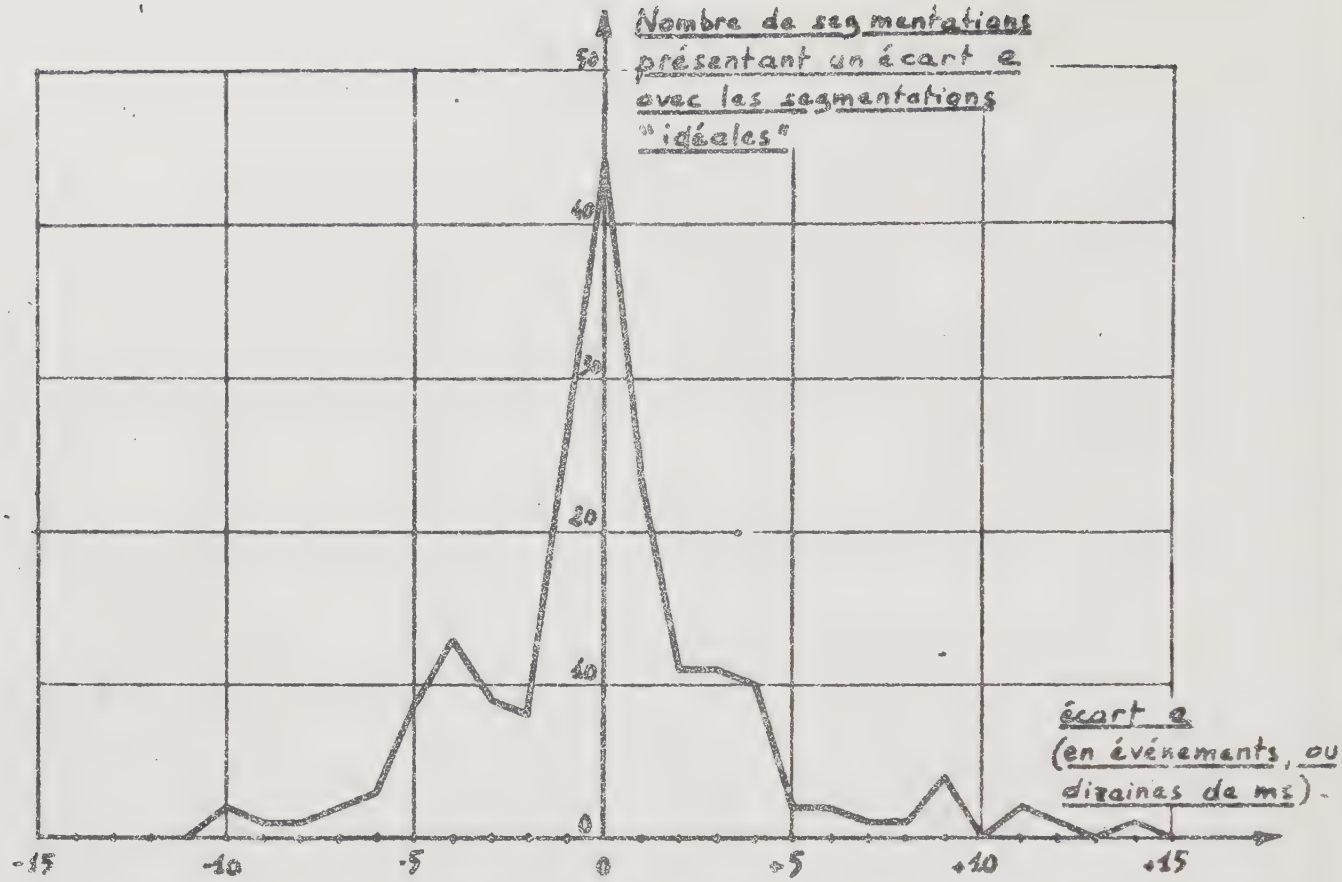
c) Fautes bénignes : décalage d'une frontière par rapport à la position "idéale", ou détection de plusieurs frontières dans un son stable. Nous avons compté 41 frontières excédentaires de ce type (non compris celles classées en b) dont la seule conséquence est de multiplier par un facteur d'environ 1,3 le nombre de phonatomes à reconnaître.

Nous avons mesuré le décalage de toutes les frontières trouvées (sauf celles signalées en b) par rapport aux positions "idéales", et construit un histogramme des écarts; ceux-ci, donnés en valeur absolue, représentent des nombres d'événements, c'est-à-dire des dizaines de ms (fig. VI-27). Pour l'interprétation de ce diagramme, on tiendra compte de ce que les positions "idéales" ne sont pas "parfaites", c'est-à-dire qu'elles sont à $\pm$ 10 ms près. On a représenté fig. VI-27 le pourcentage de frontières comprises à l'intérieur d'un écart $\pm$ e , en fonction de cet écart. On peut admettre que la

..../

Fig VI-26

Principe de la segmentation du message phonatoire, et de leur normalisation temporelle. Exemple. Les [illegible], locuteur JSL. Distance d'ordre 30 ms.



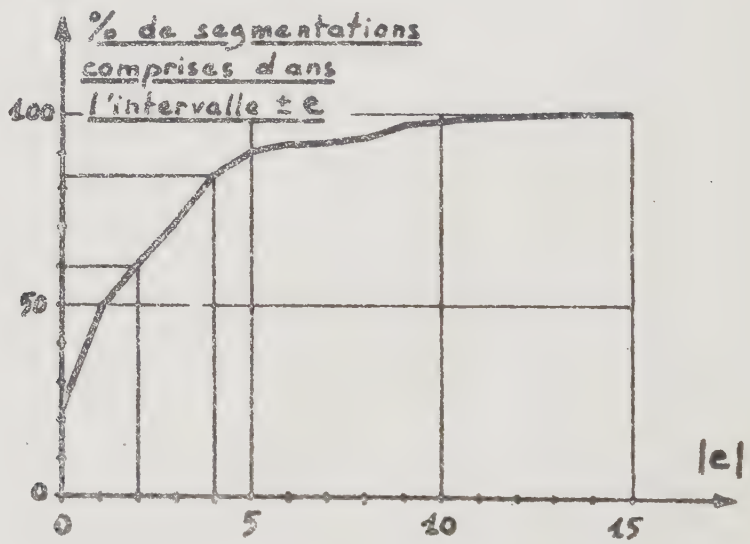


Fig VI-27

Efficacité de la segmentation. Au lieu de 145 segmentations "idéales" (frontières des phonatomes déterminées à l'examen du sonagramme, donc à ± 10 ms près), le programme a trouvé 186 segmentations. Trois segmentations idéales seulement ont été manquées. Les 44 segmentations excédentaires sont pour la plupart des segmentations multiples trouvées sur des sons stables et pourraient facilement être réduites. Tel qu'il est, le programme fournit environ 85 % de frontières "excellentes" ($|e| \leq 20$ ms) ou "correctes" ($|e| \leq 40$ ms).

segmentation est excellente lorsqu'une frontière se trouve à moins de ± 20 ms de sa position "idéale", et qu'elle est correcte jusqu'à ± 40 ms. Selon cette estimation, on aurait environ 60 % de frontières "excellentes", et environ 85 % de frontières "excellentes ou correctes". Si l'on ne conserve que la première des segmentations trouvées sur un son stable (règle facilement programmable), ces mêmes chiffres passent respectivement à 70 % et 90 %.

Nous pouvons donc estimer que la segmentation est dans l'ensemble satisfaisante, et nous passons maintenant à l'apprentissage des phonatomes.

VI.4.2. Apprentissage et reconnaissance des phonatomes

Nous avions le choix entre deux attitudes : ou bien utiliser directement comme références les phonatomes du dictionnaire de l'Icophone, ou bien utiliser une partie de notre corpus pour "apprendre" certains phonatomes, et l'autre partie pour tester la reconnaissance. Après quelques essais nous avons préféré - provisoirement - la seconde attitude, qui nous semblait plus souple.

Le programme RECO, qui effectue les opérations d'apprentissage et de reconnaissance, peut fonctionner selon trois modes différents. Si la carte de données ne comporte que le numéro de la séquence étudiée, celle-ci est simplement affichée, avec les courbes associées et le découpage en phonatomes normalisés. Pour obtenir le fonctionnement en apprentissage, l'opérateur doit donner, après le numéro de séquence, la transcription phonétique de celle-ci, correspondant au découpage exact effectué par le programme; c'est pourquoi il est nécessaire d'avoir lancé auparavant le programme en impression seule sur la même séquence, pour pouvoir affecter un nom à chaque phonatome normalisé. Le fonctionnement en reconnaissance est obtenu en disposant un caractère spécial à la suite du numéro de séquence.

VI.4.2.1. Apprentissage

Après lecture sur disque et segmentation, le programme associe à chaque phonatome normalisé un "nom", c'est-à-dire un ensemble de deux symboles phonétiques de la carte de données. Si, par exemple, la carte porte les indications :

272 .MARII

cela signifie que la séquence n° 272 comporte les phonatomes normalisés

.M, MA, AR, RI, II

Dans cet exemple le 4e phonatome rencontré reçoit le nom "RI", et un numéro lui est attribué dans une table matricielle, à l'intersection de la ligne R et de la colonne I . C'est exactement la logique de commande de l'Icophone, utilisée ici presque sans changements; en particulier la table matricielle (MATOR) est la même.

Connaissant ce numéro qui servira d'adresse, on peut mémoriser le phonatome dans le dictionnaire. Mais si l'on rencontre ultérieurement ce

...../

même phonatome, il faut pouvoir utiliser la nouvelle information. Pour cela, on relit le phonatome mémorisé, et l'on calcule les valeurs spectrales d'un nouveau phonatome-référence en faisant la moyenne point par point entre l'ancien phonatome-référence et la nouvelle occurrence. Plus généralement, si l'on rencontre la $n^{ième}$ occurrence $X_n(i,j)$ du même phonatome, on forme le nouveau phonatome-référence $R_n(i,j)$ par l'opération récurrente :

$$R_n(i,j) = \frac{1}{n}\left[(n-1)\,R_{n-1}(i,j) + X_n(i,j)\right]$$

avec $i = 1$ à 5 et $j = 1$ à 64

Il est évident que le résultat est équivalent à

$$R_n(i,j) = \frac{1}{n}\sum_{\nu=1}^{\nu=n} X_\nu(i,j)$$

On a donc à chaque instant dans $R_n(i,j)$ la moyenne de toutes les occurrences déjà rencontrées du même phonatome, à condition de conserver l'indice n avec chaque nouvelle référence. Ce processus conduit à un affinage progressif - et automatique - de chaque phonatome.

Nous avons obtenu par ce moyen un dictionnaire relatif à l'une des 8 voix enregistrées, comprenant 107 phonatomes différents, élaborés à partir de 270 occurences (voir quelques échantillons fig. VI-28). On pourrait ainsi penser que chaque phonatome-référence résulte d'une moyenne établie sur 2 ou 3 occurences; en réalité, les phonatomes stables (du type /a a/, /y y/, /s s/, etc..) ayant donné lieu à des segmentations multiples, les phonatomes plus significatifs n'ont guère été élaborés qu'à partir de 1 ou 2 occurences. On ne peut donc être certain que ces phonatomes constituent d'excellentes références.

VI.4.2.2. Reconnaissance

Le processus de reconnaissance consiste simplement à rechercher le taux de coïncidence entre le phonatome incident et chacun des phonatomes références; après quoi les résultats sont triés et les noms des 20 phonatomes-références les plus proches sont imprimés (fig. VI-29).

Nous avons, dans un premier temps, utilisé le programme d'apprentissage pour créer un dictionnaire réduit, à partir de la phrase "Les petits oiseaux chantent, c'est le printemps" prononcée par le locuteur numéro 1 avec une intonation courante. La reconnaissance a porté sur les phonatomes de la même phrase, prononcée peu après par le même locuteur avec une intonation plus grave, et monotone, donc avec des différences certaines sur le plan de la durée, du rythme et de l'accentuation.

Sur les 26 phonatomes normalisés de la seconde séquence, 22 ont été correctement reconnus, les 4 fautes étant les suivantes :

..../

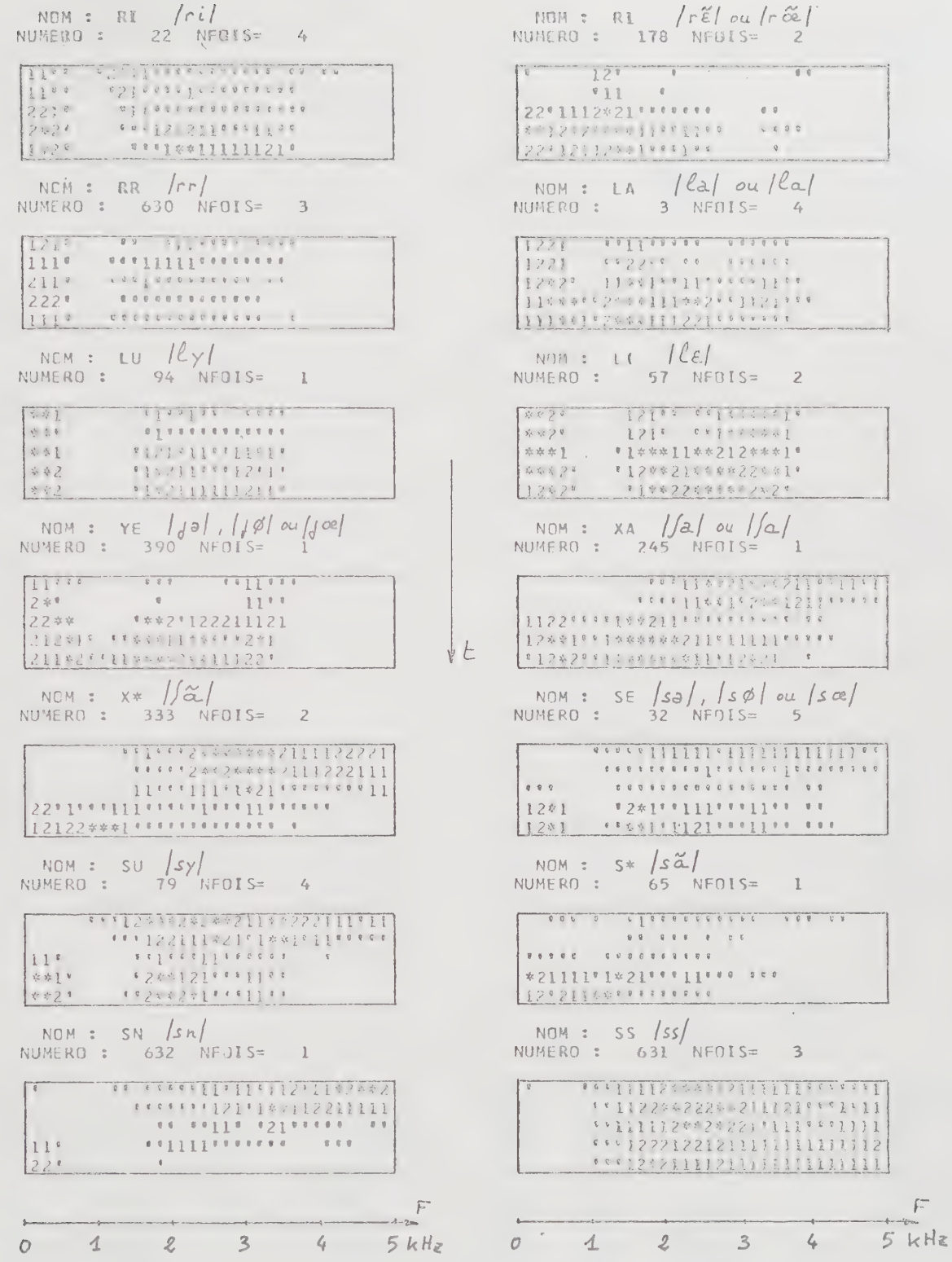


Fig VI-28

Quelques échantillons du dictionnaire de 107 phonatomes normalisés, créé par apprentissage sur une seule voix (9 phrases, locuteur JSL). Chaque phonatome représente la moyenne de plusieurs occurrences, dont le nombre est indiqué ("NFOIS").

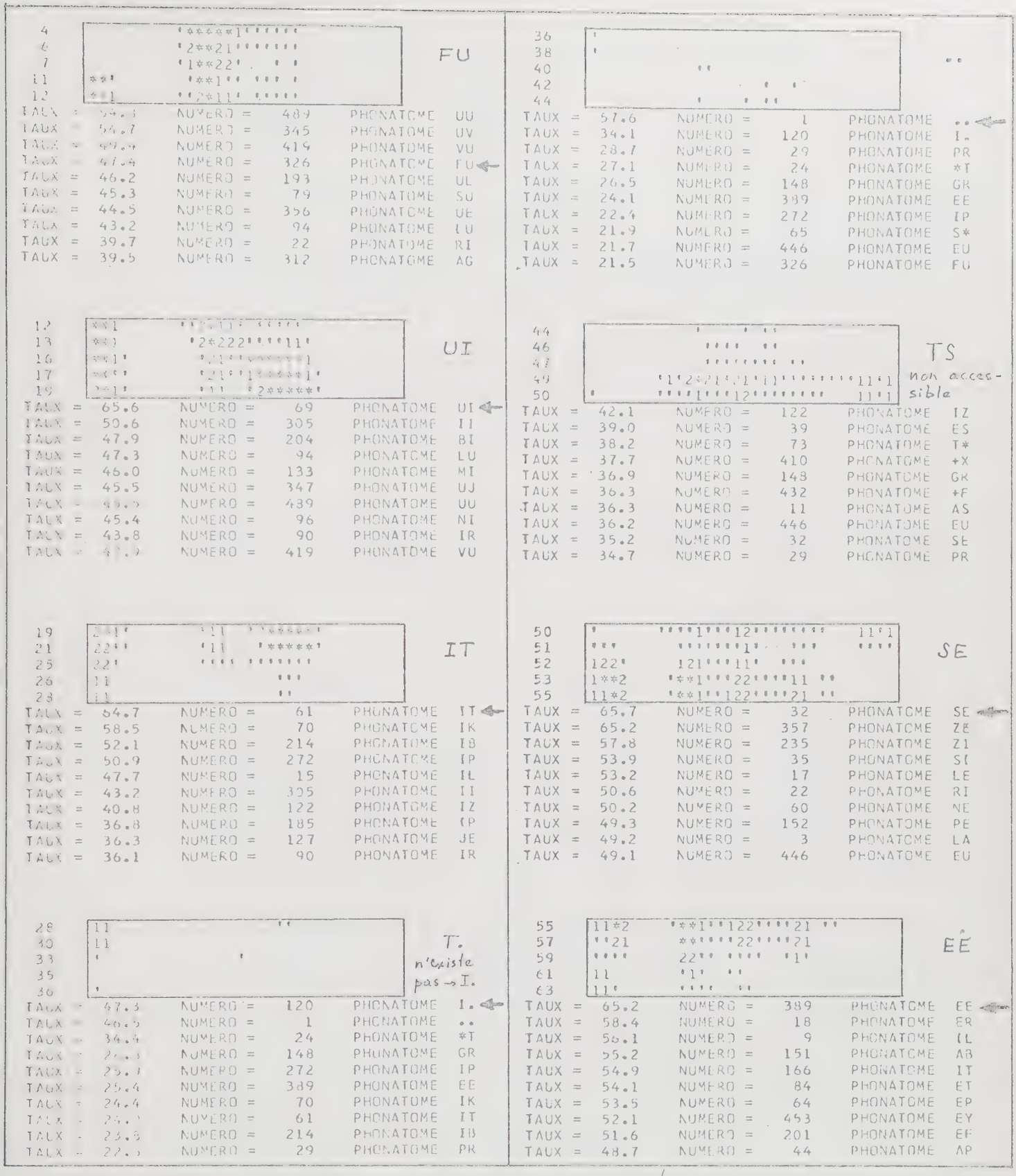


Fig VI-29

Reconnaissance de la séquence n° 84 (mot "fuite" isolé, locuteur JSL) à partir du dictionnaire de 107 phonatomes. Le phonatome-référence TS est inaccessible dans le dictionnaire par suite d'une erreur de programmation. Sur les 8 phonatomes déterminés par la segmentation, 6 sont reconnus correctement. Mais après regroupement en phonèmes on obtient la suite "UUIT..SEE". Les groupes "UU", "T..S" et"EE" se ramènent à "U", "T" et "E" par des considérations facilement programmables. La seule faute grave est donc la non-reconnaissance du "F" initial.

/œp/ au lieu de /œt/, qui arrive en 2e position

/ɑ̃ɑ̃/ " " /ɑ̃t/ " 3e "
/sɛ̃/ " " /sœ̃/ " 3e "
/œp/ " " /œ̃t/ " 4e "

Comme on le voit, ces fautes sont minimes, et tout à fait compréhensibles sur le plan phonétique, car /œ/ et /œ̃/ sont voisins, ainsi que /p/ et /t/. Nous avons décompté, pour arriver à 26 phonatomes "utiles", 8 phonatomes correspondant à des silences en fin de séquence, et 4 phonatomes qui n'étaient pas disponibles dans le dictionnaire (à cause d'une erreur de programmation découverte par la suite).

Ce résultat nous a incités à étendre l'expérience. Nous avons constitué un nouveau dictionnaire, à partir de 9 séquences (phrases complètes) du même locuteur, représentant 107 phonatomes-références différents. La reconnaissance a porté sur d'autres séquences provenant du même locuteur, séquences dont le contenu phonétique représentait des mots isolés.

Les 13 mots testés en reconnaissance donnèrent lieu à 108 phonatomes utiles (mêmes règles de décompte que précédemment). Sur cet ensemble nous avons pu relever 34 erreurs, ce qui correspond à un taux de reconnaissance de 68 % sur les phonatomes.

VI.4.2.3. Assemblage en symboles phonétiques

La suite de digrammes phonétiques n'est pas le résultat intéressant en reconnaissance automatique de la parole. Soit par exemple à reconnaître la phrase "Bonjour messieurs", dont la transcription phonétique est /bõʒurmesjø, - La décomposition en digrammes est /bõ/, /õʒ/, /ʒu/, /ur/, /rm/, /me/, /es/, /sj/, /jø/). Un système qui reconnaîtrait parfaitement tous ces digrammes donnerait en réalité deux fois trop d'information, puisque la connaissance d'un digramme sur deux suffirait à reconstituer le message.

On peut donc tolérer une efficacité relativement faible du système, si l'on prévoit une procédure permettant de tirer parti de cette redondance. De toute manière, si le système a fait une erreur sur un phonatome, il faut bien prendre une décision plus ou moins arbitraire : supposons par exemple que l'on ait obtenu la suite /bõ/ , /iʒ/ /ʒu/ , etc..., le premier symbole phonétique est nécessairement/ /;mais pour déterminer le second il faut faire un choix entre /õ/ et /i/. Si ce choix est fait intelligemment (c'est-à-dire en tenant compte des tendances du système à certaines erreurs plutôt qu'à d'autres), le taux de reconstitution en symboles phonétiques doit être supérieur au taux de reconnaissance des phonatomes.

Pour vérifier cette idée, nous avons appliqué aux résultats précédents le critère simple suivant : en cas de conflit entre deux phonatomes successifs, le phonème-frontière est choisi dans le phonatome ayant donné lieu au meilleur taux de reconnaissance. Ceci était une manière de faire confiance à la qualité acoustique du système, qui nous semblait faire des erreurs phonétiquement vraisemblables.

Les 13 mots de l'expérimentation précédente ont donné lieu à 132 pho-

...../

natomes, dont 24 ne pouvaient être reconnus correctement, faute d'avoir accès aux phonatomes-références correspondants. Cependant le système leur a attribué un nom, évidemment faux. Les 58 erreurs (34 et 24) correspondent donc à un taux global de reconnaissance de 56 % sur les phonatomes. Après assemblage de ces phonatomes on pouvait dénombrer 107 phonèmes corrects sur 145, soit 74 % de réussite. Il est donc évident que l'utilisation de ce processus d'assemblage améliore considérablement les résultats.

Malgré la portée limitée de cette expérimentation, portant sur un dictionnaire réduit (107 phonatomes au lieu de 600) et un seul locuteur, il semble que les options essentielles de nos recherches y trouvent une confirmation; d'autant que les phonatomes-références ont été établis à partir d'un apprentissage extrêmement succinct. Il faut encore souligner que l'utilisation de données préenregistrées, qui constitue par ailleurs un gros handicap (volume de traitement, absence de renouvellement du corpus), présente au moins l'avantage d'exclure totalement l'adaptation de l'homme à la machine, qui fausse souvent les résultats.

VI.4.3. Regroupement des phonèmes en mots

La suite phonétique issue d'un système de reconnaissance analytique présente en général des lacunes et des erreurs; par ailleurs, elle n'est pas segmentée en mots (sauf de manière artificielle, en ménageant une pause entre chaque mot). Il est intéressant de rechercher dans quelle mesure la connaissance préalable du vocabulaire utilisé (sous forme de lettres ou de symboles phonétiques) permet d'opérer cette nouvelle segmentation, et même éventuellement de corriger certaines erreurs de reconnaissance. Cette seconde opération doit être possible, car l'information contenue dans un mot est beaucoup plus faible que l'information contenue dans l'ensemble des symboles phonétiques qui le composent : un mot choisi parmi mille mots équiprobables requiert 10 bits, alors que les 4 symboles phonétiques qui le constituent (en moyenne), choisis parmi 30 équiprobables, correspondent à 20 bits. La redondance de la suite phonétique doit être de l'ordre de 50 %.

Nous n'avons jusqu'à présent abordé que la première opération : étant donnée une suite ininterrompue de phonèmes, issue d'une transcription phonétique manuelle (donc supposée correcte), étant donné également l'ensemble des mots utilisés, disponibles sous forme de symboles phonétiques, il s'agissait d'intercaler automatiquement des blancs entre chaque mot.

Ce programme (LAIUS, réalisé par A. BRAURE) reposait initialement sur un principe extrêmement simple : le premier caractère du texte est comparé à tous les mots du lexique; s'il ne peut constituer un mot, on lui ajoute le caractère suivant, et on consulte à nouveau le lexique. Et ainsi de suite, jusqu'à ce que l'on trouve le mot recherché. Après quoi le caractère suivant est considéré isolément, et l'on recommence une nouvelle investigation. En classant judicieusement les mots du lexique, la recherche peut être extrêmement rapide.

Pour la commodité de lecture, nous donnerons les exemples sous forme orthographiée. Les données étaient constituées par la transcription phonétique continue du texte suivant :

" Il règne une chaleur moite; on entend quelque part un léger sifflement, comme s'il y avait une fuite à la tuyauterie de chauffage ".

...../

Dans le lexique figuraient tous les mots de ce texte, et quelques autres, comme "chat", "leur", "moi", "ouate", "an", "tant", "parrain", etc... Le premier résultat fut curieux; qu'on en juge : "Il règne une chat leur mou a ton an tant quelque pas rein les jets sifflement comme si li avait une fut y tas la tuyauterie de chauffe a je ". Soit environ 13 fautes sur les 24 mots du texte original.

Pour éviter ce morcellement, nous avons adopté la règle suivante : quand une suite de symboles phonétique (limitée à 10 caractères) peut correspondre soit à plusieurs mots courts, soit à un seul mot long, ce dernier est systématiquement choisi. Sur le texte précédent, le programme n'a fait que deux erreurs : "parrain" au lieu de "par un", et "cil" au lieu de "s'il". Sur un texte plus long (300 mots) on a pu compter une quarantaine d'erreurs (ce nombre n'a pas de signification en soi, car il dépend des mots ajoutés dans le lexique).

L'efficacité de cette règle simple nous a surpris, car il semblerait de prime abord que la réduction des ambiguïtés dépende essentiellement de critères syntaxiques et sémantiques plutôt que phonétiques ou lexicaux. L'explication a peut-être pour origine la remarque suivante : plus un mot est court, plus il est fréquent et donc porteur de moins d'information (et réciproquement, les mots fréquents tendent à se raccourcir, p.ex. : cinématographe devenant cinéma, puis ciné; automobile devenant auto, etc...). L'auditeur cherche dans le discours les mots-clés, qui lui permettent de comprendre au plus vite le sens général de la phrase. Ces mots-clés sont en général les plus informatifs, c'est-à-dire les plus longs. Sachant cela implicitement, le locuteur choisit ses mots de façon à éviter les ambiguïtés : si une suite de mots courts peut être interprétée comme un seul mot long perturbant la signification de la phrase, ces mots courts sont rejetés et le locuteur en choisit d'autres. Ainsi, par une convention tacite relevant sans doute de l'économie du langage, la signification d'ensemble d'un groupe phonétique serait-elle prioritaire par rapport à la signification des sous-groupes constituants. Nous retrouvons ici la notion de forme ou de structure, avec ses qualités de cohésion (formes fortes et formes faibles) et de hiérarchie (formes et superformes).

VI.5. RECONNAISSANCE GLOBALE

La reconnaissance globale de mots isolés et répertoriés semble devoir mener à court terme à des réalisations opérationnelles. Dans cette optique, il est indispensable de prendre en compte les problèmes pratiques tenant au coût de l'appareillage et à son fonctionnement en temps réel. Ces contraintes interviennent surtout dans le volume de mémoire nécessaire et dans le temps de traitement, quantités qu'il convient de réduire au minimum, sans éliminer naturellement l'information utile. Nous pratiquerons donc en premier lieu une triple compression de l'information, portant sur la fréquence, l'intensité, et le temps. Cette expérimentation est simulée, à partir des données dont nous disposons sur disque.

Nous décrirons ensuite un processus dynamique de reconnaissance, qui revient à chercher la distance minima entre deux configurations malgré les déformations localisées de l'échelle de temps. Nous donnerons les résultats fournis par plusieurs variantes de ce programme, dont la mise au point n'est pas encore terminée.

...../

VI.5.1. Compression de l'information

VI.5.1.1. Compression fréquentielle

Les 64 voies d'analyse que nous avons obtenues par F F T sont sans doute excédentaires; nous avons pu constater que 32 fréquences suffisaient à donner de la parole un diagramme temps-fréquence significatif. S'il s'agit de suivre l'évolution d'ensemble du squelette sémantique, 16 voies doivent être suffisantes. Néanmoins, nous n'avons pas réparti ces voies linéairement profitant de la grande finesse de l'analyse F F T nous avons placé le maximum de résolution aux alentours de 1500 Hz, en utilisant le sous-programme ADELE décrit plus haut. Par exemple la voie n° 1 représente la partie du spectre inférieur à 600 Hz; la voie n° 6 représente les fréquences comprises entre 1300 et 1600 Hz, et la voie n° 16 représente les fréquences supérieures à 4200 Hz.

VI.5.1.2. Compression d'intensité.

Les valeurs spectrales étaient jusqu'à présent représentées par des nombres compris entre 0 et 1000 (10 bits). Nous avons codé chaque valeur sur 3 bits, c'est-à-dire entre 0 et 7. En effet, l'échelle en 5 nuances adoptée pour l'impression des sonagrammes numériques était à peu près suffisante pour mettre en évidence visuellement les formes sémantiques, mais seulement pour un locuteur parlant à niveau moyen. Ayant laissé de côté dans cette expérimentation le système de régulation de niveau, nous avons cru bon de conserver une marge dynamique supplémentaire. Dans une expérimentation portant sur le signal réel, un bon compresseur dynamique et quelques précautions de prise de son permettraient sans doute de coder les amplitudes spectrales sur 2 bits (intervalle 0-3) voire un seul.

VI.5.1.3. Compression temporelle

La compression temporelle est plus délicate : l'intervalle de 10 ms est déjà presque trop grand par rapport à certains phénomènes très brefs, fréquents dans la parole. Il n'est donc pas question de l'augmenter ou de prélever régulièrement un évènement sur deux ou trois, sous peine d'une détérioration rapide de l'information à traiter.

Nous avons utilisé ici encore la courbe de stabilité temporelle S(t). Après avoir vérifié que la courbe obtenue était acoustiquement correcte malgré les compressions de fréquence et d'intensité, nous avons prélevé dans la configuration X (i,j) les évènements les plus significatifs, correspondant aux extrema de la courbe, c'est-à-dire aux frontières et aux pivots des phonatomes.

Soit Y (k,j) la nouvelle configuration formée par assemblage de ces événements représentatifs. Une compression supplémentaire permet d'éliminer tout événement ressemblant trop au précédent (problème des segmentations excédentaires évoqué plus haut) : l'événement $Y_{k+1}(j)$ est comparé à l'événement $Y_k(j)$. Si le taux de coïncidence est supérieur à un cer-

...../

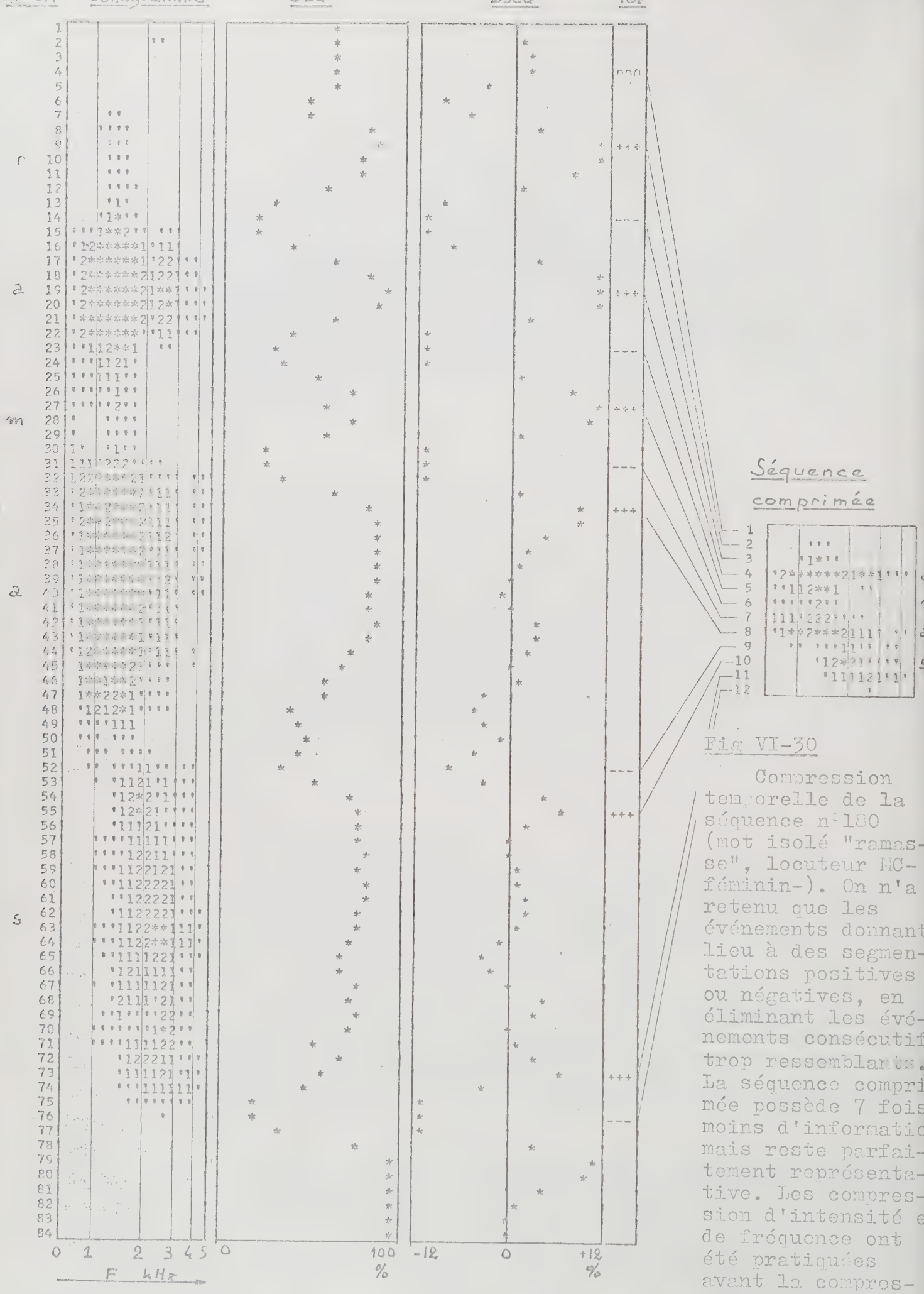


Fig VI-30

Compression temporelle de la séquence n°180 (mot isolé "ramasse", locuteur MC-féminin-). On n'a retenu que les événements donnant lieu à des segmentations positives ou négatives, en éliminant les événements consécutifs trop ressemblants. La séquence comprimée possède 7 fois moins d'information mais reste parfaitement représentative. Les compression d'intensité et de fréquence ont été pratiquées avant la compression temporelle.

tain seuil (65 % dans la version actuelle du programme), l'événement $Y_{k+1}(j)$ est éliminé, et on compare $Y_k(j)$ à $Y_{k+2}(j)$, etc, jusqu'à trouver pour $Y_{k+n}(j)$ un taux de coïncidence inférieur au seuil; à ce moment les événements $Y_k(j)$ et $Y_{k+n}(j)$ deviennent respectivement les événements $Z_\ell(j)$ et $Z_{\ell+1}(j)$ de la version comprimée $Z(\ell,j)$ de la séquence $X(i,j)$

La compression ainsi définie est très efficace (fig. VI-30). Elle permet de ne conserver en moyenne qu'un événement sur 5 environ. De plus, la longueur de la séquence comprimée devient une grandeur significative du point de vue phonétique, car elle ne dépend plus (du moins plus autant) de la rapidité d'élocution. Si la séquence est décrite phonétiquement par N symboles (blancs de début et de fin compris), on trouve environ 2 N événements dans la séquence comprimée.

L'ensemble des compressions effectuées sur la fréquence, l'intensité et le temps permet de diviser par 70 environ le volume d'information à traiter, par rapport aux données issues de l'analyse. Si l'on adopte un débit de parole moyen (2 mots de 5 phonèmes par seconde), le débit d'information après compression est de l'ordre de $20 \times 16 \times 3 \simeq 1\,000$ bits par seconde. La reconnaissance proprement dite consiste à passer de 1 000 à 50 bits par seconde.

Nous n'avons pas tenté de pousser plus loin la compression; il nous semblait plus logique de mettre d'abord au point un processus de reconnaissance, et de traiter ultérieurement ce problème en utilisant comme critère objectif le taux de reconnaissance.

VI.5.2. Mémorisation et reconnaissance

VI.5.2.1. Conditions d'apprentissage

Parmi les données acoustiques enregistrées, nous ne disposons pas de mots répétés plusieurs fois par le même locuteur. Nous avons, par contre, des mots et phrases prononcés par les huit locuteurs ("petits fichiers"). Nous avons mémorisé comme références 7 phrases et 7 mots prononcés une seule fois par l'un d'entre eux (locuteur masculin). Les phrases sont indiquées au § IV.2.1.2.c; les mots sont "hippophage", "fascicule", "birmanie", "ramasse", "cerise", "bizarre", "babar". Les séquences-références sont ainsi voisines, puisque les mots précédents sont extraits des phrases. Les séquences à reconnaître sont les mots et phrases analogues, prononcés par les 7 autres locuteurs (4 féminins, 3 masculins). Ces conditions sont donc infiniment plus sévères que celles des expérimentations usuelles en reconnaissance vocale, qui portent généralement sur le même locuteur, avec apprentissage plus ou moins prolongé.

VI.5.2.2. Principe de la reconnaissance

Le programme que nous allons décrire a été réalisé en juillet 1971, donc avant que nous ayions connaissance (à BUDAPEST, août 71) des méthodes

...../

de programmation dynamique utilisées par SAKOE et CHIBA (bib. VI-20).

Soit à comparer deux séquences $X(i,j)$ et $R(k,j)$ de même contenu phonétique, mais présentant entre elles des différences temporelles (anamorphose temporelle irrégulière). Les durées totales sont représentées par N_x et N_R. Pour simplifier les notations nous appellerons X_i et R_k les événements correspondant aux instants i et k.

Le programme ZAT compare initialement X_1 et R_1. Si le taux de coïncidence C est inférieur à un seuil C_o, X_1 est comparé à R_2, puis R_3 etc.., jusqu'à l'événement R_m pour lequel $C(X_1, R_m) \geqslant C_o$. X_2 est ensuite comparé à R_{m+1}, puis si nécessaire R_{m+2} etc.. Si, au-delà de l'événement R_{m+M}, on n'a pas trouvé d'événement R_k donnant un taux de coïncidence suffisant, on passe à l'événement X_3, que l'on compare à R_{m+1} et ainsi de suite. Pratiquement, nous avons fixé M = 10; la "fenêtre de coïncidence temporelle" est donc de 10 événements comprimés, soit environ 0,5 seconde de temps réel. On établit ainsi une série de correspondances entre événements, matérialisées,sur un exemple (fig. VI-31). Ces correspondances illustrent une transformation temporelle discrète $k = f(i)$ pour laquelle la coïncidence entre les configurations $X(i,j)$ et $R[f(i),j]$ est maxima, moyennant les limitations posées (largeur M de la fenêtre de coïncidence, valeur du seuil C_o).

Le taux global de coïncidence est égal à la somme des taux partiels, rapportée à la longueur de la plus longue des deux séquences. Soit

$$C(X,R) = \frac{\sum_{i=1}^{i=N_x} C\left[X_i, R_{f(i)}\right]}{\max\left[N_x, N_r\right]}$$

Cette opération est effectuée pour chacune des références mémorisées. Il suffit ensuite de déterminer la référence pour laquelle on a obtenu un taux maximum.

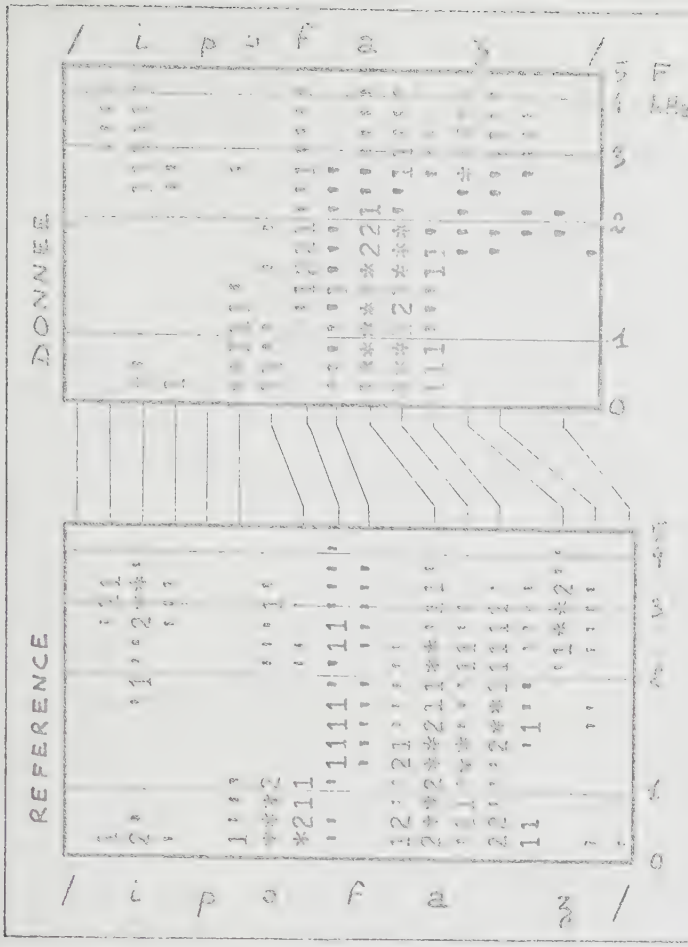
L'ensemble de la reconnaissance ainsi conçue peut se dérouler en temps réel : il n'est point besoin d'attendre que la séquence X soit terminée pour commencer le traitement, et l'algorithme est extrêmement rapide. Le temps de traitement est à peu près proportionnel au nombre de séquences en mémoire.

VI.5.2.3. Résultats

Avec le programme précédent nous avons obtenu au 1er passage 47 reconnaissances rectes sur 96, soit un taux de 49 %. Ce taux varie selon les locuteurs entre 36 % et 70 %. Ce résultat peut être considéré comme bon, étant donné les conditions d'apprentissage et le fait qu'aucune anamorphose fréquentielle n'a été pratiquée.

Nous avons défini un autre indice caractérisant la qualité de la reconnaissance, le rang moyen. Si la séquence à reconnaître arrive en première position, elle est "reconnue" et son rang est 1. Si elle arrive en 4e position, elle n'est pas reconnue, et son rang est 4. Le rang moyen, moyenne des

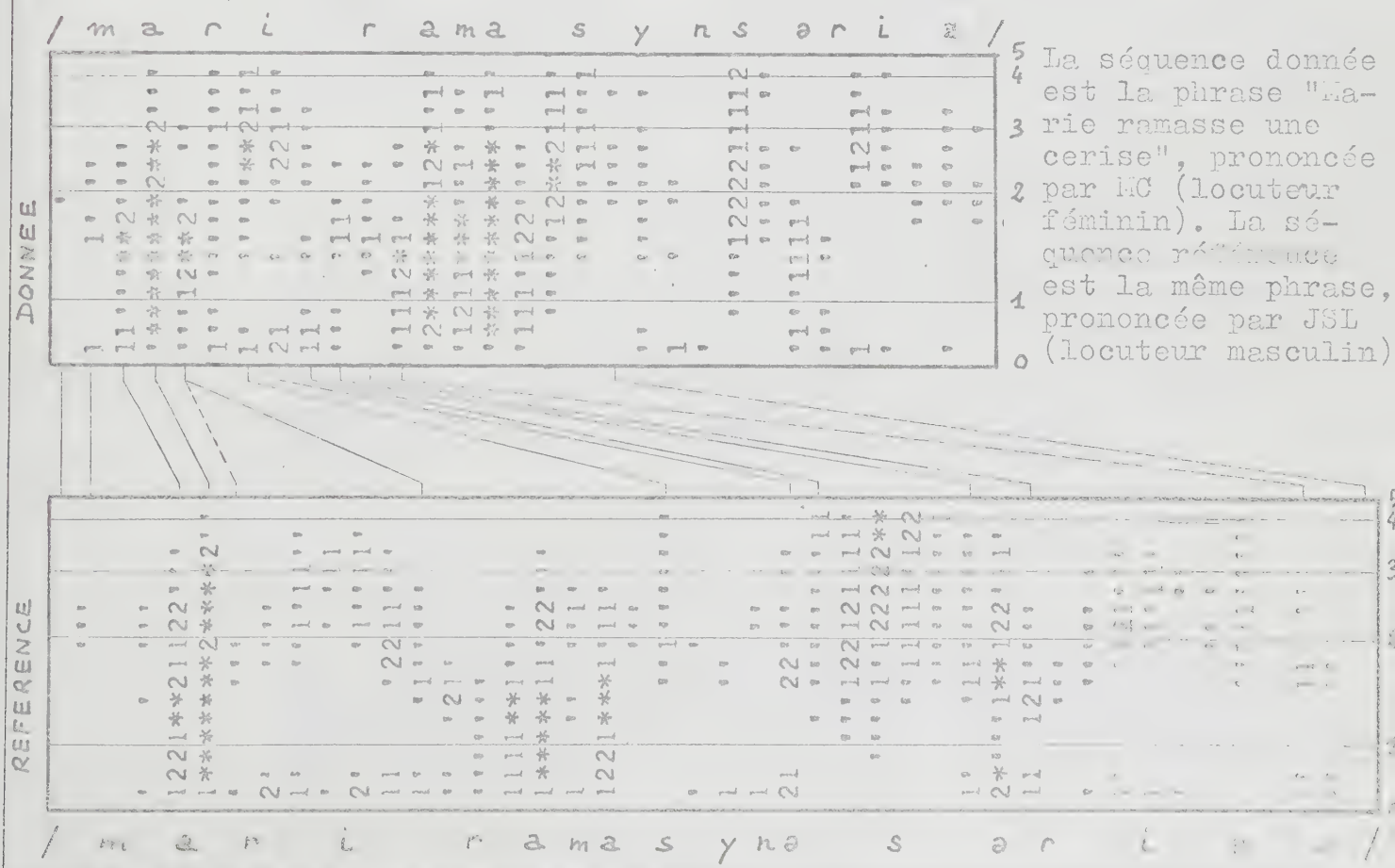
...../



Reconnaissance globale [illegible] méthode des corr[illegible] gramation dynamique). La [illegible] n° 216 (mot "[illegible]" [illegible] féminin FR) est reconnue à [illegible] de la séquence n° 76 (même [illegible] locuteur masculin JSL). Les [illegible] ces sont comprimées en temps [illegible] quence et intensité. Il existe entre les deux séquences une anamorphose temporelle telle que leur simple juxtaposition donnerait un taux de coïncidence très faible. Une correspondance entre deux événements est établie lorsque leur taux de coïncidence est $\geqslant 40\ \%$; le taux moyen pour l'ensemble de la séquence passe ainsi à 50 %, et permet sa reconnaissance parmi les 14 mots-références.

Fig VI-32

Erreur typique commise par le programme ZAT. La correspondance indiquée en pointillé a été manquée ; les correspondances suivantes sont erronées du point de vue phonétique, et le taux global de coïncidence est très faible.



La séquence donnée est la phrase "Marie ramasse une cerise", prononcée par MC (locuteur féminin). La séquence référence est la même phrase, prononcée par JSL (locuteur masculin)

rangs obtenus pour toutes les séquences étudiées, caractérise non pas la réussite, mais les erreurs du système : si la séquence à reconnaître est classée 2e, nous estimons que le résultat est malgré tout meilleur que si cette séquence était classée 10e. Cette amélioration se reflète dans le rang moyen, alors qu'elle passe inaperçue dans le taux de reconnaissance. Naturellement, à un taux de reconnaissance de 100 % correspond un rang moyen unité.

Dans l'expérimentation précédente le rang moyen obtenu est de 2,8, ce qui n'est pas extraordinaire : rappelons qu'un choix aléatoire fournirait un taux de reconnaissance de 7 % et un rang moyen voisin de 7; comme les mots et les phrases ont des longueurs très différentes un choix aléatoire tenant seulement compte de ces longueurs donnerait un taux de 14 % et un rang moyen de 3,5. Donc notre système fait des erreurs bien caractérisées; schématiquement, si une séquence n'est pas reconnue, elle semble presque rejetée en fin de classement.

Ces erreurs sont facilement explicables : lorsqu'une correspondance entre événements X_i et R_k a été manquée (soit à cause d'une mauvaise prononciation, soit parce que le seuil C_o n'est pas atteint), le programme va la chercher plus loin. S'il la trouve, en dépassant la correspondance suivante, un décalage temporel apparaît de manière irrémédiable, puisqu'aucune possibilité de retour n'est prévue (fig. VI-32). Les coefficients M et C_o ont donné lieu à une optimisation ultérieure : actuellement (avril 1972) le taux de reconnaissance obtenu avec ZAT, dans les mêmes conditions que précédemment, est de 70 % (avec M = 10 événements et C_o = 44 %).

VI.5.2.4. Variantes

Le programme TAZ en cours de mise au point, tente de remédier au défaut de ZAT : la fenêtre temporelle de coïncidence s'étend de part et d'autre de la correspondance précédente; la correspondance est établie sur l'événement fournissant un maximum de coïncidence, à l'intérieur de la fenêtre temporelle. De plus, les séquences comparées R et X sont normalisées à un même niveau moyen, et la plus petite des deux est utilisée comme base de comparaison (ces deux opérations ne sont pas dans l'optique du temps réel).

Tous ces essais n'ont pas encore permis une amélioration substantielle des résultats : si certains locuteurs sont reconnus à 80 %, le taux moyen de reconnnaissance n'excède pas 50 %, et le rang moyen ne descend guère au-dessous de 2 . Cependant, certains points positifs sont apparus : les séquences reconnues ne sont pas les mêmes que précédemment; une reconnaissance quasi-totale devrait donc être possible. D'autre part le gain sur le rang moyen indique que les erreurs sont plus cohérentes. Enfin les résultats reflètent la parenté plus ou moins grande entre les voix étudiées et la voix utilisée comme référence : les taux obtenus sont en corrélation avec les coefficients d'anamorphose tels que le programme ADAPT les a déterminés.

VI.5.2.5. Efficacité pour un locuteur unique

Comme nous l'avons vu, les données préenregistrées dont nous disposons ne permettent pas de tester le programme sur un seul locuteur. Nous avons cependant de bonnes raisons de penser que l'efficacité est très grande dans ce cas.

...../

Au cours de la mise au point du programme ZAT, les mêmes séquences ont été utilisées en mémorisation et en reconnaissance, mais avec de grandes différences dans la compression temporelle. Nous comparions donc l'une à l'autre deux versions comprimées d'une même séquence, versions composées d'événements différents. Les 14 séquences étudiées de cette manière ont donné lieu à une reconnaissance totale.

Ceci n'est évidemment pas un résultat valable, mais seulement une indication de la souplesse du programme dans le domaine temporel.

VI.6 DEVELOPPEMENTS ET PERSPECTIVES

Nous indiquerons ici les principales directions de recherche qui se dégagent des expérimentations décrites. Une amélioration commune à tous les programmes réside évidemment dans un appareillage analogique d'analyse de la parole connecté directement au calculateur,qui permettra d'introduire la parole en machine en temps réel, de renouveler le corpus phonétique, d'étudier l'adaptation au locuteur et d'utiliser des traitements interactifs : locuteur, système de reconnaissance, synthétiseur, auditeur, etc... Cet analyseur-codeur est en cours de réalisation; aussi n'en dirons-nous pas plus dans le cadre du présent travail Cependant, certains points peuvent être approfondis par simulation, tant en reconnaissance analytique que globale. Nous détaillerons quelque peu ces points; et montrerons que les deux approches sont tout à fait convergentes. Nous évoquerons enfin quelques problèmes liés à un niveau supérieur de la reconnaissance, problèmes sémantiques d'organisation des connaissances dans le cerveau et de polysensorialité.

VI.6.1. Développements de la reconnaissance analytique

Les principales voies à explorer sont au nombre de trois : utilisation comme référence du dictionnaire de phonatomes normalisés mis au point pour la synthèse, mise au point d'une recherche plus rapide des phonatomes-références, et étude de l'assemblage en phonèmes, puis en mots.

VI.6.1.1. Utilisation du dictionnaire de synthèse

Cette option est importante : elle revient à supprimer complètement l'apprentissage, c'est-à-dire à mettre au point un processus efficace d'adaptation au locuteur. Ce processus devra inclure l'anamorphose fréquentielle régulière, et sans doute une anamorphose irrégulière, ainsi qu'une normalisation dynamique et spectrale. De plus, le dictionnaire de synthèse doit subir encore certaines retouches, et doit être adapté au format utilisé en reconnaissance : 5 événements de 64 ou 32 voies couvrant la bande 0-5000 Hz, au lieu de 20 événements en 44 voies couvrant la bande 0-4400 Hz. Ce dernier point a été réalisé récemment (M. MLOUKA). Un exemple des phonatomes obtenus est visible sur la figure V-15.

...../

VI.6.1.2. Recherche rapide en mémoire

C'est le point essentiel d'une optimisation du programme : celui-ci doit devenir plus maniable (modulaire) et adaptable au traitement en temps réel. Il est tout à fait superflu de comparer un phonatome incident à tous les phonatomes-références. Le volume de calcul sera considérablement réduit en faisant un pré-classement des phonatomes incidents, soit selon les spectres stables correspondant aux frontières des phonatomes, soit selon la répartition grossière de l'énergie dans diverses zones du phonatome. On pourra également faire intervenir dans la reconnaissance les statistiques exposées au chapitre IV, qui amélioreront le résultat d'ensemble.

VI.6.1.3. Assemblage de la suite de phonatomes

On peut tout d'abord envisager la transmission des adresses des phonatomes reconnus. La logique de l'Icophone placé en bout de ligne serait alors extrêmement simple, et le débit d'information sur la ligne de transmission serait de l'ordre de 100 bits par seconde au lieu de 50; de plus, cette méthode sauvegarderait vraisemblablement le maximum d'information acoustique, dans la mesure où les phonatomes reconnus de manière erronée sont voisins des phonatomes "idéaux".

L'assemblage des symboles phonétiques pourrait être effectué de manière moins rudimentaire que nous ne l'avons fait. Il faudrait approfondir l'étude des erreurs phonétiques du système, et adopter un processus d'assemblage qui compense statistiquement ces erreurs.

Enfin le regroupement en mots, indispensable dans certaines applications (commande vocale, dialogue avec les calculateurs), permet d'assimiler le système à un système de reconnaissance globale avec, cependant, un avantage considérable : les mots sont alors stockés en mémoire sous la forme peu encombrante de symboles phonétiques, et le lexique peut être pratiquement illimité. De plus, cette approche nous semble la plus apte à une généralisation, dans le cas de la parole courante, dont le débit est continu et le vocabulaire étendu.

VI.6.2. Développements de la reconnaissance globale

Dans ce domaine les voies à explorer nous semblent en relation avec la possibilité d'une réalisation technologique à court terme. Ce sont : la mise au point de l'algorithme dont nous avons donné dans ZAT une bonne ébauche, la reconnaissance continue d'un vocabulaire limité, et surtout l'utilisation de mots-références formés à la demande, à partir du dictionnaire et de la logique de l'Icophone.

VI.6.2.1. Mise au point de la méthode de correspondance temporelle

Il serait bon, tout d'abord, de pouvoir tester le programme sur un seul locuteur et un vocabulaire plus vaste, afin de comparer ses performan-

...../

ces à celles des systèmes décrits dans la littérature. Il faudrait ensuite améliorer la méthode des correspondances en autorisant certains retours en arrière, qui pourraient d'ailleurs être pénalisés par une diminution arbitraire du taux de coïncidence. Enfin, il serait nécessaire d'appliquer, comme en reconnaissance analytique, les anamorphoses (régulière, irrégulière) et régulations (dynamique, spectrale) permettant une meilleure adaptation au locuteur.

VI.6.2.2. Application à la parole continue

Le principe que nous avons adopté dans ZAT est transposable au cas de la parole continue, à condition de tenir compte des problèmes exposés lors de l'assemblage des symboles phonétiques en mots (règle du mot le plus long). L'algorithme serait alors sensiblement celui de LAIUS, et ne concernerait plus des symboles phonétiques, mais des événements représentatifs. Il est évident que la reconnaissance ne pourrait porter que sur un vocabulaire limité (sans doute à quelques dizaines ou centaines de mots, pour rester dans les limites d'un calculateur moyen), et se ferait avec un délai de principe (donc incompressible) de l'ordre de 0,5 seconde. La compréhension humaine nécessite d'ailleurs souvent un délai de cet ordre.

VI.6.2.3. Utilisation du dictionnaire de synthèse

On peut imaginer que les mots-références soient composés au moyen de la logique de l'Icophone; il suffirait alors de frapper sur le clavier leur contenu phonétique, ce qui supprimerait tout apprentissage, du moins sous forme vocale. Les mots ainsi reconnus seraient transmis sous forme d'un simple numéro de lexique, moyennant un débit de l'ordre de 20 bits par seconde, à un Icophone en bout de ligne ou à un système de commande vocale. L'intérêt ne serait pas tant la diminution du débit que la possibilité d'utiliser sans apprentissage un vocabulaire ajustable pour chaque application, et modifiable instantanément. C'est là sans doute une voie de recherches susceptible de mener rapidement à une réalisation technologique valable.

VI.6.3. Les problèmes sémantiques

Nous nous sommes limités, dans cette étude, aux aspects acoustique et phonétique du message parlé. Cependant, le fond du problème de la communication parlée se trouve au niveau sémantique : l'objet du langage est la transmission d'idées et non de mots. Dans ce sens, le message parlé ne saurait être isolé des autres messages sensoriels, présents ou passés, reçus et intégrés par l'auditeur. Nous allons tenter d'éclaircir quelque peu certains aspects du problème de la compréhension, préalable indispensable à leur éventuelle mécanisation. Nous avons formulé l'essentiel de ces idées, ainsi que les grandes options adoptées ultérieurement au sujet de la reconnaissance automatique, en février 1969, lors d'une réurion de notre groupe de recherche (bib. VI-25). Nous espérons nous faire pardonner le caractère spéculatif de ce paragraphe, en rappelant qu'il ne vise qu'à définir des directions de recherche, et en mettant l'accent sur ses motivations et ses buts pratiques.

...../

VI.6.3.1. Quelques définitions

Les définitions qui suivent n'ont rien d'exhaustif : ce sont seulement les acceptions que nous donnerons dans la suite à certains termes utiles

a) Concept. Nous désignons par ce terme la représentation mentale unique d'un ou plusieurs faits d'expérience.

Le "fait d'expérience" est un élément ("objet", ou "phénomène") du monde extérieur (espace-temps), appréhendé par l'intermédiaire de nos sens, et auquel nous pouvons attribuer une certaine unité fonctionnelle.

Un concept intègre donc les informations fournies par les différents sens.

Si beaucoup de mots ("signifiants", que nous avons appelés plus haut des mots-idées) désignent des concepts ("signifiés"), il existe nombre de concepts qui ne se laissent pas facilement définir par un seul mot, de même que certains mots (ceux que nous avons appelés des mots-chevilles) ne désignent pas des concepts, mais marquent la syntaxe de la phrase.

b) Idée. Une idée est une association de plusieurs concepts, liés par des relations plus ou moins complexes. L'ensemble des relations internes d'une idée constitue sa structure, qui possède une existence autonome et obéit à certaines lois.

La "phrase" (au sens large) constitue une représentation concrète de l'idée; ce n'est pas la seule possible : tout comportement volontaire, passant ou non par la verbalisation écrite ou orale, permet de matérialiser une idée par une action.

c) Pensée. Nous englobons dans ce terme l'ensemble des opérations mentales, conscientes et inconscientes : formation des concepts, formation et association des idées, transposition des idées à des concepts différents, etc. Nous supposons de plus que toute pensée est motivée, à plus ou moins long terme, de manière plus ou moins consciente, par une action à produire sur le monde extérieur.

Sans doute ces définitions et hypothèses sont-elles sommaires; elles permettent cependant de souligner l'aspect individuel de la communication. Concepts, idées, pensée, sont spécifiques à chaque individu, et la communication ne peut jamais être qu'approximative.

VI.6.3.2. Les trois niveaux de compréhension de la parole

a) Niveau I (acoustique-phonétique)

Ce niveau correspond à la traduction du message parlé en une suite de symboles phonétiques; il requiert la connaissance du code phonétique utilisé (dictionnaire des phonatomes)

⟹ / k o m ã t a l e v u /

...../

b) Niveau II (phonétique-linguistique)

A ce niveau est effectuée la transformation de la suite phonétique en une suite de mots, éventuellement orthographiés (version écrite). Cette opération implique la connaissance des mots utilisés (avec leur prononciation) et des règles syntaxiques (et éventuellement grammaticales) de la langue

/ k o m ɑ̃ t a l e v u / ⟹ " Comment allez-vous ?"

c) Niveau III (linguistique-sémantique ou conceptuel).

Le passage du linguistique au conceptuel nécessite une "culture" (ensemble de concepts) commune aux deux interlocuteurs, ainsi que la connaissance des principales structures idéelles repérées par la syntaxe. C'est à ce niveau que se situe la compréhension d'un message tel que

" 'jour, M'sieur, 't-allez-vous ? "

L'idée transmise dans ce message est rudimentaire. Son "information perçue", ou imprévisibilité (cf. chap. II) est totalement déterminée par le contexte : elle est quasi-nulle, lorsque la phrase est prononcée par un monsieur qui en rencontre un autre, et une simple poignée de main suffit à transmettre la même idée. L'information perceptive est plus importante, si la phrase est lâchée au milieu d'une conversation, ou lorsqu'un monsieur rencontre une dame. Plus le contexte rend l'idée prévisible, plus la quantité d'information physique nécessaire est faible.

Le contexte joue donc un rôle déterminant dans la communication, rôle d'autant plus difficile à définir qu'il est, lui aussi, individualisé. Le monde extérieur fournit en permanence une énorme quantité d'information physique par l'intermédiaire des divers canaux sensoriels; mais chaque individu n'y puise que l'information perceptive nécessaire à l'évolution de sa propre pensée.

d) Relation entre les trois niveaux

La communication se fait dans l'ordre inverse de celui que nous avons adopté pour l'exposé. Le contexte (qui peut être simplement le passé du message) permet généralement de déterminer à l'avance les concepts ou idée dont il va être question. Le message peut être très incomplet, ou déformé; s'il subsiste seulement quelques repères acoustiques (fragments de squelette sémantique, intonation, etc...) la compréhension n'en est pas affectée. Il peut être nécessaire de passer au niveau II, c'est-à-dire au décodage des mots et de leurs relations syntaxiques, si le contexte est insuffisant (message imprévu). Le recours au niveau I n'est que rarement pratiqué dans la conversation courante, car il est coûteux et peu sûr (nécessit de segmentation, d'identification des sons un par un, etc...); il n'intervient guère que lorsque le contexte est contradictoire, ou le mot inconnu (noms propres, langues étrangères, etc...)

..../

VI.6.3.3. Acquisition et organisation des concepts

Le répertoire des concepts est pour chaque individu en constante évolution, dans son volume et dans sa structure. L'acquisition et l'organisation des concepts ne sont pas des problèmes indépendants.

a) Organisation des concepts

Tous les concepts du répertoire sont liés les uns aux autres de multiple manière. La nature des relations existant entre deux concepts peut reposer sur la similitude, ou l'antinomie de certaines des images sensorielles qui les composent, sur la proximité dans le temps ou dans l'espace des faits d'expérience correspondants, sur l'utilisation conjointe des concepts dans une action, etc...

Il semble que chaque concept soit affecté d'une distance vis-à-vis de chacun des autres; tout concept peut jouer le rôle de centre d'intérêt pour l'ensemble du répertoire.

b) Acquisition des concepts

La distance entre deux concepts n'est pas figée : elle évolue avec l'apprentissage. Ce que nous appelons éducation est un apprentissage des concepts, dirigé par l'extérieur (la société) et rationalisé : l'éducation présente les concepts de façon à favoriser certaines relations, qui par la suite feront surgir certaines idées plus facilement que d'autres.

L'éducation n'est pas tout l'apprentissage : l'homme fabrique seul son expérience. Cet auto-apprentissage suit un processus d'essais et d'erreurs : un essai est enregistré comme erreur lorsque son résultat va à l'encontre ce certains principes "vitaux" (survie, conservation de l'espèce, etc...)

IV.6.3.4. La compréhension automatique de la parole

La solution ultime du problème de reconnaissance vocale est surbordonnée à la réalisation d'un système capable d'apprendre et d'utiliser des concepts. Ce système doit rendre compte des fonctions et structures qui viennent d'être évoquées. Sa mise au point nous semble devoir se faire en plusieurs étapes.

a) Utilisation des formes

Chez l'homme les données sensorielles sont des Formes. La réponse d'une cellule (visuelle, auditive, etc..) n'a pas d'importance en soi; l'information élémentaire est constituée par la réponse d'ensemble des cellules, qui constitue une Forme, avec toutes ses propriétés (talité, anamorphose, hiérarchie).

Il importe donc d'apprendre à utiliser les Formes, c'est-à-dire à les discerner, les déformer, les assembler, les reconnaître. Les anamorphoses et normalisations que nous avons appliquées à la parole constituent une première ébauche de ce travail, et il faudrait généraliser à plusieurs

...../

dimensions des traitements analogues.

La structure séquentielle des calculateurs actuels nous semble très gênante pour ces opérations, car très éloignée des structures neuroniques examinées au chapitre III. La simulation sur machine séquentielle d'un réseau de cellules interconnectées est extrêmement lourde et limitée.

b) Passage des mots aux concepts

Un même concept, et a fortiori une même idée, peuvent être exprimés par une multitude de mots et de phrases différents. Il serait fructueux d'étudier la structuration du répertoire de concepts telle que divers mots ou phrases revêtent une même signification, en commençant au besoin par un langage réduit. Cette étude pourrait utiliser un apprentissage dirigé.

c) Autoapprentissage

Pour passer à l'autoapprentissage, il faut doter le système d'un "principe vital" et d'un moyen d'expression, ainsi que d'un organe lui permettant de recueillir la réaction du monde extérieur. Le système doit prendre l'initiative d'un "comportement" vis-à-vis du monde extérieur; estimer lui-même la valeur "vitale" de son essai et en tirer la leçon, qui se traduit par une modification des relations entre ses concepts.

d) Etude du contexte et de la prévisibilité

Ces problèmes peuvent être abordés avec un système capable d'autoapprentissage. Après chaque mot du message, le système propose un nouveau mot, choisi parmi ceux qui lui semblent les plus vraisemblables. Si le mot suivant coïncide avec le mot proposé, c'est que les critères de choix étaient bons; sinon on tire la leçon de l'échec, on ajuste une nouvelle stratégie et on passe au mot suivant.

Il est bien entendu que de telles recherches, dans la mesure où elles sont ambitieuses, risquent d'être fort décevantes. Cependant leur application à la reconnaissance de la parole leur éviterait d'être un but en soi, et leur efficacité pourrait être jaugée au moyen de résultats pratiques et chiffrés.

o
o o

VII - CONCLUSION

Au terme de cette étude nous allons tenter de dresser le bilan des principaux problèmes abordés. Certains d'entre eux n'ont d'ailleurs été qu'effleurés, bien que nous ayions parfaitement conscience de leur importance fondamentale dans la communication parlée : mécanismes perceptifs, problèmes sémantiques, par exemple. A vrai dire, ces questions ne peuvent actuellement que faire l'objet d'hypothèses, et nous avons préféré nous limiter aux domaines acoustique et phonétique, dans lesquels la méthode expérimentale - même simulée - conduit à des résultats concrets.

Les méthodes physiques d'étude du signal, pour lesquelles la parole n'est qu'un signal acoustique parmi d'autres, se heurtent dès l'abord aux exigences de la perception : la transformation de FOURIER doit s'adapter à la notion de spectre instantané, et y parvient très mal, puisqu'elle ne peut rendre compte de la sensibilité de l'oreille en fréquence et en temps; la théorie de l'information perd toute rigueur lorsque le récepteur est intelligent et sait utiliser le contexte ou découvrir la structure du message. Cependant la transposition visuelle de la parole obtenue par analyse temps-fréquence permet, moyennant un certain entraînement, de faire correspondre les sensations auditives à des documents objectifs.

L'audition et la perception auditive sont au coeur du problème; mais l'anatomie et la physiologie se noient rapidement dans la complexité des réseaux neuroniques. Les seules indications positives que nous possédions sont obtenues de manière externe, en étudiant les réponses à des stimuli bien définis. Cette méthode suppose un morcellement de la perception, contre lequel réagissent la théorie de la Forme et le Structuralisme. La notion de Forme, au sens de totalité perceptive, est pour la suite notre fil directeur, comme la représentation temps-fréquence est notre outil.

Ainsi armés, nous pouvons déterminer les structures acoustiques et phonétiques de la parole en commençant par l'étude de l'appareil phonatoire. Les multiples possibilités de celui-ci, en tant que générateur sonore, laissent entrevoir un invariant, le squelette sémantique, qui présente tous les caractères d'une Forme : caractères de totalité, d'anamorphose, de hiérarchie.

La notion d'anamorphose fréquentielle est établie à partir de l'étude des voyelles, et la possibilité de décomposition d'une Forme est mise à profit pour définir des éléments nommés phonatomes, ou diphonèmes, qui, à la différence des phonèmes, respectent la continuité de la parole. Leur répartition dans la langue obéit à des lois statistiques, et reflète l'existence de diverses contraintes phonétiques. La question trouve un prolongement dans la fabrication de "mots aléatoires", qui respectent statistiquement ces contraintes.

La synthèse de la parole montre la valeur pratique des idées précédentes. La constitution d'un dictionnaire de phonatomes normalisés oblige à maîtriser leur structure, et le synthétiseur permet d'étudier expérimentalement l'anamorphose fréquentielle. En outre, la synthèse possède un débouché technologique important dans le domaine de l'informatique.

La reconnaissance automatique de la parole est un sujet difficile, et passionnant dans la mesure où il n'est pas encore résolu. Après avoir examiné ses principaux aspects et décrit quelques réalisations-types, nous avons effectué par simulation un prétraitement de la parole, déduit de l'expérience acquise en analys

...../

et en synthèse, qui vise à mettre en évidence le squelette sémantique, quel que soit le locuteur. Nous avons ensuite proposé deux voies différentes pour la reconnaissance : l'une, analytique, consiste à isoler et normaliser les phonatomes, puis à les comparer aux phonatomes-références stockés par apprentissage; l'autre, globale, dirigée vers la possibilité de réalisation à court terme, considère des mots isolés, comparés dans leur ensemble à des mots-références.

Nous n'avons pas cherché à obtenir les meilleurs performances, ni à optimiser les programmes, mais à dégager des principes. Les résultats obtenus, en particulier dans la normalisation fréquentielle et temporelle de la parole, nous semblent suffisamment démonstratifs pour justifier la poursuite de telles recherches. Leur objectif n'est pas seulement technologique; replacé dans le cadre sémantique, un système de reconnaissance automatique constitue un véritable modèle de la perception humaine.

ANNEXE I

NOTATION PHONETIQUE

L'alphabet phonétique international utilise 36 signes phonétiques pour transcrire la langue française. Dans une première étude nous n'en avons distingué que 29 (30 si l'on y inclut le silence, représenté par un point), estimant peu significatives les distinctions entre /œ̃/ et /ɛ̃/ , /ɑ/ et /ɑ̃/, /œ/, /ə/ et /ø/, /ɲ/ et /nj/, /y/ et /ɥ/, /u/ et /w/. En effet, ces phonèmes sont couramment intervertis ou confondus, sans que la compréhension de la parole en soit altérée.

Pour représenter ces symboles nous avons dû nous contenter des caractères usuels en informatique : lettres majuscules et caractères spéciaux. Il était, de plus, nécessaire de faire correspondre un symbole phonétique à un seul caractère-machine, sous peine d'alourdir considérablement la programmation.

Les transcriptions phonétiques sont toujours placées entre barres obliques. Ce caractère étant utilisé par ailleurs comme symbole phonétique, les textes phonétiques en code-machine sont généralement placés entre guillemets.

Voici les équivalences entre les symboles des deux systèmes, avec une série de mots-clés illustrant leur prononciation. Nous avons adopté, dans la mesure du possible, un classement des phonèmes selon leur parenté acoustique (de l'aigu au grave dans chaque catégorie; alternance sourde-sonore pour les consonnes.

...../

	Code international	code-machine	mot-clé
Voyelles orales	/i/	I	lit
	/e/	)	été
	/ɛ/	(	marais
	/y/	U	Ursule
	/œ/)		(peur)
	/ə/) (	E	(petit)
	/ø/)		(jeu
	/a/)		(patte
	/ɑ/) (	A	) (pâte
	/ɔ/	O	sol
	/o/	+	saule
	/u/	W	bijou
Voyelles nasales	/ɛ̃/)		(brin
	/œ̃/) (	1	) (brun
	/ɑ̃/	*	chant
	/ɔ̃/	/	bonjour
Semi-voyelles	/j/	Y	paille
	/ɥ/		lui
	/w/		Louis
Consonnes fricatives	/s/	S	saucisson
	/z/	Z	zèbre
	/ʃ/	X	chat
	/ʒ/	J	janvier
	/f/	F	fameux
	/v/	V	vert

..../

Consonnes nasales	/n/	N	Nantes
	/m/	M	menthe
	/ɲ/		agneau
Consonnes liquides	/l/	L	salon
	/r/	R	bureau
Consonnes plosives	/p/	P	pari
	/b/	B	barbare
	/t/	T	bateau
	/d/	D	badaud
	/k/	K	écart
	/g/	G	langue

ANNEXE II

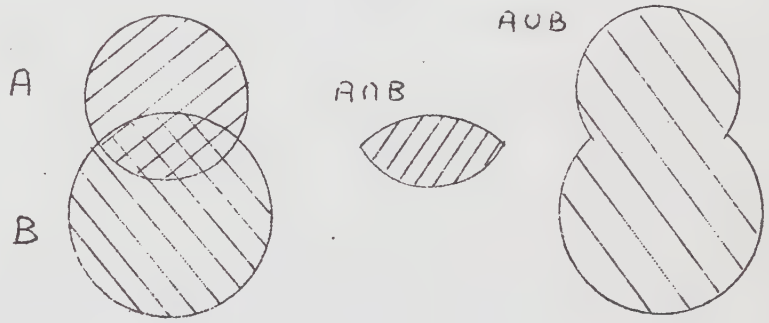
TAUX DE COINCIDENCE ENTRE DEUX CONFIGURATIONS

Pour toutes les comparaisons d'évènements, de phonatomes ou de squelettes sémantiques effectués dans cette étude, nous avons adopté une définition particulière de la distance. C'est le taux de coïncidence, ou rapport de la partie commune au tout.

1 - FORME BINAIRE

Cette forme est dûe à D. TEIL, qui l'a utilisée pour comparer les uns aux autres les phonatomes du dictionnaire de synthèse. Le taux de coïncidence entre deux configurations binaires A et B est le rapport

$$\tau_o(A,B) = \frac{A \cap B}{A \cup B}$$



PROPRIETES

- τ_o est nul dès que A et B sont disjointes, ce qui n'est pas le cas d'une distance.
- τ_o n'est égal à 1 que si A et B sont identiques.
- τ_o est indéterminé si A et B sont nulles toutes les deux.
- L'opération peut être étendue à un nombre quelconque de configurations. Pour trois configurations A, B et C on définit

$$\tau_o(A, B, C) = \frac{A \cap B \cap C}{A \cup B \cup C}$$

2 - ADJONCTION D'UN SEUIL

Dans la plupart des applications, on considère deux configurations nulles comme identiques. Ceci est obtenu en formant

$$\tau_K(A, B) = \frac{K + A \cap B}{K + A \cup B}$$

...../

La constante K permet de régler le seuil au-dessous duquel les configurations A et B sont considérées comme nulles : on voit aisément que τ_k (A,B) reste voisin de 1 tant que A U B reste très inférieur à K. En reconnaissance de parole le seuil K permet d'éviter les perturbations dûes au bruit de fond.

3 - FORME GENERALE

Soient les configurations A (i,j) et B (i,j) avec i = 1, 2 ...., M et j = 1, 2 ..., N - Leur taux de coïncidence est donné par l'expression

$$\tau_k\ (A,B) = \frac{K + \sum_{i=1}^{M} \sum_{j=1}^{N} \min \left[A\ (i,j)\ ,\ B\ (i,j) \right]}{K + \sum_{i=1}^{M} \sum_{j=1}^{N} \max \left[A\ (i,j)\ ,\ B\ (i,j) \right]}$$

Cette expression est valable quelle que soit la forme numérique de A et B - En notation binaire, si l'on admet que

$$\min\ (0,1) = 0$$

$$\text{et}\quad \max\ (0,1) = 1,$$

on retrouve l'expression exposée en 2).

L'opération peut être aisément étendue à un nombre quelconque de dimensions et de configurations; par exemple pour trois configurations A, B, C à trois dimensions, on aura :

$$\tau_k(A,B,C) = \frac{K + \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{P} \min \left[A\ (i,j,k)\ ,\ B\ (i,j,k)\ ,\ C\ (i,j,k) \right]}{K + \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{P} \max \left[A\ (i,j,k)\ ,\ B\ (i,j,k)\ ,\ C\ (i,j,k) \right]}$$

On cherche généralement à établir un taux de coïncidence entre deux configurations. Dans ce cas, la programmation est extrêmement simple (sous-programme COMPA) et le traitement est très rapide.

B I B L I O G R A P H I E

Abréviations :

- G A M - Groupe d'Acoustique Musicale de l'Université PARIS VI.
- J A S A - Journal of the Acoustical Society of America.

I.1. - R. SIESTRUNCK, E. LEIPP - Quelques méthodes et problèmes de l'acoustique - Annales de l'Université de Paris, n° 2 - 1968.

I.2. - E. LEIPP - Acoustique et musique. Ed. MASSON, Paris, 1971

II.1. - J.S. LIENARD - Analyse et synthèse de FOURIER d'un son permanent. Edition interne. C.N.A.M. 1965.

II.2. - J.C. RISSET - A propos de la notion de phase. Communication non publiée. 1971.

II.3. - L. PIMONOW - Vibrations en régime transitoire. DUNOD. 1962.

II.4. - B.P. LATHI - An introduction to Random Signals and Communication Theory-International Textbook Company. Scranton, Pennsylvania, 1968.

II.5. - J.S. BENDAT, A.G. PIERSOL - Measurement and analyses of random data. John WILEY and SONS. New-York, London, Sydney, 1966.

II.6. - J.L. FLANAGAN - Speech Analysis, Synthesis and Perception.SPRINGER Verlag - Berlin 1965.

II.7. - R.B. BLACKMAN, J.W. TUCKEY - The measurement of power spectra. DOVER publications, New-York, 1959.

II.8. - W.T. COCHRAN, J. W. COOLEY and Al. (G.A.E. Subcommitte on Measurement Concepts). What is the Fast Fourier Transform. IEEE transactions, volume AU-15, n° 2, juin 1967.

II.9. - J. CONNES - Computing problems in FOURIER Spectroscopy. Centre Interdiscipline Régional de Calcul Electronique, C.N.R.S. Orsay - 1969.

II.10. - R.K. POTTER, G.A. KOPP, H.C. GREEN - Visible speech - Van NOSTRAND, New-York, 1947.

..../

II.11. - A. MOLES - Théorie de l'Information et Perception Esthétique - Flammarion, 1958.

II.12 - A. MOLES, B. VALLANCIEN - Communications et langages - (articles de M. AUBRY, G. LEHMANN, A. MOLES, P. GUIRAUD, W. MEYER-EPPLER, W.A. ROSENBLITH, A. FESSARD, R.G. BUSNEL, A. CLASS, S. BOREL-MAISONNY, D.B. FRY, G. PERDONCINI, J.C. LAFON, B. VALLANCIEN) - Gauthier-Villars, Paris, 1963.

II.13 - L. PIMONOW - Bio-informatique de l'Audition, volumes I et II. Cours à l'Ecole Pratique des Hautes Etudes, Paris 1969.

III.1. - S.S. STEVENS and H. DAVIS - Hearing, its Psychology and Physiology J. WILEY and sons, New-York, London, 1938 - 5e édition, 1960.

III.2. - H. FLETCHER - Speech and Hearing in Communications - Van NOSTRAND, New-York, 1953.

III.3. - E. LEIPP - Réflexions sur la mécanique et l'acoustique de l'oreille moyenne. Bulletin du GAM n° 49, juin 1970.

III.4. - E. LEIPP - Mesures physiques et perception des sons. Conférences de Journées d'étude du Festival du Son. Ed. CHIRON, Paris 1970.

III.5. - E. LEIPP - L'enregistrement et la reproduction des percussions. Conférences des Journées d'étude du Festival du Son. Ed. CHIRON, Paris 1965.

III.6. - J.C. RISSET - Sur l'analyse, la synthèse et la perception des sons, étudiées à l'aide de calculateurs électroniques - Thèse de doctorat ès-sciences physiques - Faculté des Sciences d'Orsay, 1967.

III.7. - J. BARBIZET - Etudes sur la mémoire. Ed. L'expansion Scientifique française - PARIS 1966.

III.8. - E. LEIPP - Le problème de l'intelligibilité de la parole. Bulletin du GAM n° 37 - 1968.

III.9. - J. MONOD - Le hasard et la nécessité. Edition du Seuil, PARIS 1970.

III.10. - W. KOHLER - Gestaltpsychology - Ed. LIVERIGHT, New-York 1929 - Edition française : Psychologie de la Forme, Gallimard 1964 (collection Idées).

III.11. - P. GUILLAUME - La psychologie de la Forme - Flammarion 1937.

..../

III.12. - P. NAVILLE - La psychologie du Comportement, Gallimard 1963 (collection Idées) - Première édition en 1942.

III.13. - E. LEIPP - Information sémantique et parole : essai d'une Gestalt-theorie - Bulletin GAM N° 22, juin 1966.

III.14. - E. LEIPP - Le problème de la perception des signaux acoustiques par effet de contraste : les signaux d'avertissement. Annales des Télécommunications, t. 20, n° 5-6, 1965.

III.15. - J. PIAGET - Le structuralisme, Presses Universitaires de France, Paris 1968 - (collection Que Sais-je).

III.16. - A. MOLES, B. VALLANCIEN - Phonétique et Phonation (articles de M. ARSLAN, L. CROATTO, D.B. FRY, B. MALMBERG, O. METTAS, G. MOLINARI, G. PERDONCINI, J. PERELLO, A. RETHI, A. SONNINEN, G. YANNATOS). Ed. MASSON, Paris 1966.

III.17. - A. MOLES - Art et Ordinateur - Ed. CASTERMAN 1971.

III.18. - A. MOLES - Méthode cybernétique et structures linguistiques - in référence III-16, ci-dessus.

III.19. - G.A. MILLER and J.C.R. LICKLIDER - The Intelligibility of interrupted Speech - JASA - 22 pp 167-173, 1950.

III.20. - W.A. ROSENBLITH - La perception catégorielle des phénomènes sonores. in réf. II.2. ci-dessus.

III.21. - A. MOLES - Aspect temporel de l'intégration du message linguistique et modèles de l'audition - in. réf. II.2. ci-dessus.

IV.1. - E. LEIPP - Mécanique et acoustique de l'appareil phonatoire. Bulletin GAM n° 32, décembre 1967.

IV.2. - A. MARTINET - Eléments de linguistique générale. Collection Armand COLIN, n° 349 - 6e édition, 1966.

IV.3. - M. CASTELLENGO - Instruments de musique électroniques. Comptes-rendus des journées d'étude du Festival du son - Ed. Chiron, Paris 1967.

IV.4. - T. TARNOCZI und J. RADNAI - Eine Möglickheit automatischer Erkennung von Vokalen. Actes du 7e Congrès International d'Acoustique, Budapest 1971.

...../

IV.5. - S. BOREL-MAISONNY - Ontogénie du langage - in réf. II.2 ci-dessus.

IV.6. - F.S. COOPER, P. DELATTRE, A.M. LIBERMAN, JM. BORST, L.J. GERSTMAN : Some experiments on the perception of synthetic speech sounds. JASA 24, PP 597-606, 1952.

IV.7. - D. MOREL - J.F. AUFFRET - Etude expérimentale de la voix et de la parole du Perroquet et du Mainate. Mémoire pour l'obtention du Certificat de Capacité en Orthophonie. Faculté de Médecine de Paris, 1969.

IV.8. - E. LEIPP - Le problème du bruit. Bulletin du GAM N° 20 - 1966.

IV.9. - A. MOLES - Les musiques expérimentales. Edition du Cercle d'Art Contemporain, Paris 1960.

IV.10. - J.S. LIENARD - Les haut-parleurs. Bulletin du GAM N° 26, 1967.

IV.11. - R. COUNORD - Théorie et applications des circuits différentiateurs et intégrateurs. Thèse de doctorat ès-sciences physiques. Faculté des Sciences de Paris, 1955.

IV.12. - E. LEIPP, J.S. LIENARD, M. CASTELLENGO, J. SAPALY, D. TEIL, A. CALINE M. MLOUKA - Colloque sur la parole. Bulletin du GAM, N° 53, 1971.

IV.13. - J.S. LIENARD - Parole et ordinateur. in. réf. IV.12 ci-dessus.

IV.14. - E. LEIPP, M. CASTELLENGO, J. SAPALY, J.S. LIENARD - Structure physiqu et contenu sémantique de la parole. Colloque sur la parole organisé par le GALF, Grenoble, avril 1967 - La Revue d'Acoustique n° 3-4, 1968.

IV.15. - M. CASTELLENGO - La synthèse de la parole à l'Icophone; les problèmes de perception d'une voix synthétique. in. réf. IV.14 ci-dessus.

IV.16. - M. CASTELLENGO - Elaboration du dictionnaire des phonatomes et amélio rations apportées à la voix de l'Icophone, in réf. IV-12 ci-dessus.

IV.17. - P. GUIRAUD - Problèmes et Méthodes de la statistique linguistique. Presses Universitaires de France, Paris 1960.

IV.18. - J.S. LIENARD - Le dictionnaire des éléments phonétiques et ses applications à la linguistique. Bulletin du GAM N° 22 bis, juin 1966.

IV.19. - J.P. HATON et M. LAMOTTE - Etude statistique des phonèmes et diphonèmes dans le français parlé. Revue d'Acoustique n° 16, décembre 1971

...../

IV.20. - J.P. TUBACH - Etude des contraintes statistiques des groupements phonématiques. Centre d'Etudes pour la Traduction Automatique, Grenoble, 1970.

V.1. - J.S. LIENARD - La synthèse de la parole, historique et réalisations actuelles. Nucleus, t.10, N° 3, 1969. Publié également dans la Revue d'Acoustique, volume 3, 1970.

V.2. - DE MILLY, LE RAY, LAVOISIER, LAPLACE, FERRIEU et VICQ D'AZIR. Compte-rendu de l'Académie des Sciences, Paris, 3 Septembre 1783.

V.3. - W. VON KEMPELEN - Le mécanisme de la Parole, suivi d'une description de la machine parlante - Vienne 1791.

V.4. - J.S. LIENARD - La machine parlante de Kempelen. Bulletin du GAM N° 34, mars 1968.

V.5. - L.R. RABINER and H. LEVITT - New results in speech synthesis by rules. Comptes-rendus du 6e Congrès International d'Acoustique, Tokyo, août 1968

V.6. - J. PONCIN - Etude d'un système de synthèse de messages vocaux. Annales des Télécommunications, t. 25, n° 11-12, novembre-décembre 1970.

V.7. - E. LEIPP, M. CASTELLENGO, J.S. LIENARD - La synthèse de la parole à partir de digrammes phonétiques. Comptes-rendus du 6e Congrès International d'Acoustique, Tokyo, août 1968

V.8. - N.R. DIXON, H.D. MAXEY - Terminal analog synthesis of continuous speech using the diphone method of segment assembly. IEEE trans. audio, vol AU-16 n° 1, mars 1968.

V.9. - S. SAITO, S. HASHIMOTO - Speech syntheses system based on interphonème transition unit. Comptes-rendus du 6e Congrès International d'Acoustique, Tokyo, août 1968.

V.10. - J. QUINIO, D. TEIL - La synthèse de la parole par ordinateur à partir de digrammes phonétiques - Revue d'Acoustique t. 3, n° 9, 1970.

V.11. - D. TEIL - Etude de génération synthétique de parole à l'aide d'un ordinateur. - Thèse C.N.A.M. Paris 1969.

V.12. - J. SAPALY - Les Icophones à commandes optique et numérique. in réf. IV.12 ci-dessus.

V.13. - J. SAPALY - L'Icophone considéré comme périphérique de calculateur. Comptes-rendus du 7e Congrès International d'Acoustique - Budapest, Août 1971.

...../

V.14. - A. CALINET - Description de deux programmes pour la correction et l'anamorphose des éléments phonétiques.
in réf. IV.12 ci-dessus.

V.15. - B. MALMBERG, D.B. FRY, R. LANCIA, R. CARRE, M. ROSSI, J. PAILLE, H. LANE - Nouvelles perspectives en phonétique.
Presses Universitaires de Bruxelles, 1970.

V.16. - J. VAISSIERE - Contribution à la synthèse par règles du français.
Thèse de 3° cycle - Centre d'Etude pour la Traduction Automatique, Grenoble 1971.

V.17. - J.S. LIENARD et D. TEIL - Les éléments phonétiques et la traduction automatique du message écrit en message parlé.
Revue "Automatisme", n° 10, Octobre 1970.

VI.1. - N. LINDGREN - Directions for speech Research.
IEEE Spectrum, mars 1968.

VI.2. - J.L. FLANAGAN - Focal points in speech communication research.
Comptes-rendus du 7e Congrès International d'Acoustique, août 1971.

VI.3. - JM. KOMATITSCH - La Reconnaissance des Formes.
2e thèse (Docteur-Ingénieur) - Paris, mars 1970.

VI.4. - J.C. SIMON, C. ROCHE - Application of questionnaire theory to pattern recognition - Institut de Programmation. Université Paris VI - 1971.

VI.5. - A. DEWEZE - Traitement de l'information linguistique. Dunod - Paris 196

VI.6. - J.Y. GRESSER, G. MERCIER - Exemple de reconnaissance automatique de la parole. Commutation et Electronique N° 32, janvier 1971.

VI.7. - J.A. DREYFUS-GRAF - Machines obéissant à la parole.
Comptes-rendus du 7e Congrès International d'Acoustique.
Budapest, août 1971.

VI.8. - P. VICENS - Aspects of speech recognition by computers.
Ph. D. Thesis - Stanford University, 1969.

VI.9. - M. LINDGREN - Machine Recognition of human language.
IEEE Spectrum, t. 2, mars et avril 1965.

VI.10. - J.Y. GRESSER - Reconnaissance de mots et de langage parlé .
Journées d'Etude sur la parole. Groupe Communication parlée du GALF, Grenoble, février 1970.

VI.11. - S.P. SMITH - A Phoneme Detector.
JASA, 23, pp 446-451 - 1951.

...../

VI.12. - J.A. DREYFUS-GRAF - Phonétographe et cybernétique.
Comptes-rendus du 5e Congrès International d'Acoustique, Liège 1965.

VI.13. - D.B. FRY and P.B. DENES - The solution of some fundamental problems in mechanical speech recognition.
Language and Speech n° 1, 1958.

VI.14. - H.F. OLSON and H. BELAR - Syllable analyzer, coder and synthesizer for the transmission of speech.
I.R.E. transactions on Audio, vol. A.U. 10, n° 1, pp. 11-17, 1962.

VI.15. - M.J. VIGNERON, M. LAMOTTE, J.P. HATON et J. BREMONT.
Recherches actuelles sur l'extraction de caractéristiques et la reconnaissance de la voix parlée.
Automatisme, t. XV, n° 12, 1970.

VI.16. - J.P. TUBACH - Reconnaissance automatique de la parole.
Thèse de doctorat ès-sciences, Université de Grenoble, 1970.

VI.17. - K.H. DAVIS, R. BIDDULPH and S. BALESHEK - Automatic recognition of spoken digits.
JASA, t 24, pp 637-642, 1952.

VI.18. - P.B. DENES and M.V. MATTHEWS - Spoken digits recognition using time-frequency pattern matching.
JASA, t 32, pp 1450-1455 - 1960.

VI.19. - J. W. GLENN and M.H. HITCHCOCK - " With a speech pattern classifier, computer listens to his master's voice ".
Electronics, may 10, 1971.

VI.20. - H. SAKOE and S. CHIBA - A dynamic programming approach to continuous speech recognition - Comptes-rendus du 7e Congrès International d'Acoustique, Budapest, août 1971.

VI.21. - M. MLOUKA - Codage et analyse spectrale de la parole en vue de la reconnaissance automatique - in. réf. IV.12 ci-dessus.

VI.22. - J.S. LIENARD et M. MLOUKA - Un aspect du prétraitement de la parole en v vue de sa reconnaissance automatique.
Comptes-rendus du 7è Congrès International d'Acoustique, Budapest, août 1971.

VI.23. - G.M. MARCOUX - Conception et Réalisation d'un analyseur de signaux acoustiques complexes. Thèse de Docteur-Ingénieur, Faculté des Sciences de PARIS, Juin 1971.

VI.24. - E. LEIPP, M. CASTELLENGO, J. SAPALY et J.S. LIENARD - Les anamorphoses de la parole et leur simulation à l'ordinateur. Comptes-rendus du 7e Congrès International d'Acoustique, Budapest 1971.

VI.25. - J.S. LIENARD - Quelques réflexions à propos d'une recherche sur la reconnaissance automatique des formes du langage parlé. Communication interne, Laboratoire d'Acoustique de l'Université Paris VI, février 1969.

TABLE DES MATIERES

Pages

AVANT-PROPOS 1

I - INTRODUCTION 4

I.1. GENERALITES 4

I.2. PLAN DE L'ETUDE 5

II- ANALYSE SPECTRALE ET INFORMATIONNELLE EN ACOUSTIQUE 7

II.1. TRANSFORMATION DE FOURIER 7

II.1.1. Bases mathématiques de l'analyse spectrale 7

1. Série de FOURIER 7
2. Intégrale de FOURIER 9
3. Le théorème de convolution 11
4. Autocorrélation et spectre de puissance 11

II.1.2. Notion de spectre instantané 12

1. Définition 12
2. Choix de la fenêtre temporelle 12
3. Analyse au moyen de filtres 13
4. Loi de KUPFMULLER 14

II.1.3. Problèmes liés au traitement numérique 15

1. Théorème des échantillonnages 15
2. Transformée discrète de FOURIER 15
3. Transformée rapide de FOURIER 17

II.2. LA REPRESENTATION TEMPS-FREQUENCE 18

II.2.1. Les trois dimensions d'un signal sonore 18

II.2.2. Le sonagraphe 19

II.2.3. Représentation temps-fréquence de quelques sons élementaires 20

II.2.4. Représentation temps-fréquences de la parole 21

...../

II.3. INFORMATION, CODAGE ET REDONDANCE 21

II.3.1. Mesure de l'information 21

II.3.2. Efficacité du codage - Redondance 23

II.3.3. Cas où les symboles ne sont pas indépendants 24

II.4. CANAL, BRUIT, MESSAGES CONTINUS 25

II.4.1. Canal et sûreté de transmission 25

II.4.2. Cas des messages continus 26

II.5. INFORMATION ET COMMUNICATION ENTRE INDIVIDUS 27

II.5.1. Intersection des répertoires 27

II.5.2. Débit informatif du message parlé 27

1. Information physique 27
2. Information perceptive 28
3. Information phonétique 28
4. Information transportée par les mots 28

II.5.3. Information sémantique et information esthétique 29

II.5.4. Hiérarchie des niveaux d'information 30

1. Gain produit par la transmission des mots 30
2. Contraintes syntaxiques 31

II.5.5. Influence du contexte et de la structuration sémantique du message 31

II.5.6. Limites de la théorie de l'Information 32

III- AUDITION ET PERCEPTION AUDITIVE 34

III.1. ANATOMIE ET PHYSIOLOGIE DE L'AUDITION 34

III.1.1. Description de l'ensemble 34

III.1.2. Oreille moyenne et adaptation ossiculaire 36

III.1.3. Le codage cochléaire 37

III.2. QUELQUES ASPECTS DE LA PERCEPTION AUDITIVE 39

III.2.1. Perception de l'intensité 39

a. Sensibilité différentielle et décibels 39
b. Paradoxe de l'intensité subjective fractionnée ... 39
c. Hypothèse de l'adaptation retardée 40
d. Perception du niveau absolu 40
f. Perception de l'intensité en fonction de la fréquence 40

III.2.2. Perception de la fréquence 41

a. Seuils différentiels et échelle logarithmique 41
b. Zone sensible 41
c. Perception fréquentielle des sons complexes 42

III.2.3. Perception du temps - Mémoire 42

a. La constante de temps de l'oreille 42
b. Constante de temps, bande passante et mémoire immédiate 43
c. Différents types de mémoire 44

III.3. FORMES, STRUCTURES ET CATEGORIES 44

III.3.1. La théorie de la Forme 44

III.3.2. Lois des Formes 45

a. Non-additivité des parties 45
b. Transposabilité et anamorphose 45
c. Emergence d'une Forme à partir d'éléments .....
d. Loi de la "bonne Forme" 45
e. Formes et superformes 46
f. Formes fortes et formes faibles 46
g. Figure et fond 46

III.3.3. Théorie de la Forme et Structuralisme 47

a. Le structuralisme 47
b. Ordre proche et ordre lointain 47

III.3.4. La perception catégorielle 48

...../

III.4 MODELE FONCTIONNEL DE L'AUDITION 49

III.5 CONCLUSION 50

IV - STRUCTURE PHONETIQUE ET ACOUSTIQUE DE LA PAROLE 51

IV.1. LA PHONATION 51

IV.1.1. Description sommaire du système phonatoire 51

1. Le larynx 52
2. Le conduit vocal 52

IV.1.2. Fonctionnement de l'appareil phonatoire 53

1. Fonctionnement du larynx 53
2. Fonctionnement du conduit vocal 54

a. Transmission acoustique 54
b. Génération de bruits et d'explosions 55

IV.2. STRUCTURE PHONETIQUE DE LA PAROLE 56

IV.2.1. Quelques définitions et remarques 56

1. La double articulation 56
2. Le point d'articulation 56
3. Phonème et contexte 56
4. Phonème et position d'articulation 57
5. Notation phonétique 57

IV.2.2. Les voyelles 58

1. Les voyelles françaises 58
2. Les fréquences des formants 59
3. Le triangle des formants 59
4. Les formants des voyelles dans la parole courante .. 60
5. Les traits distinctifs des voyelles 61

IV.2.3. Les consonnes 61

...../

IV.3. STRUCTURE ACOUSTIQUE DE LA PAROLE 62

IV.3.1. Caractéristiques externes du message parlé 63

1. Bandes passante 63
2. Dynamique 63
3. Distorsions 64

a. Altérations de la courbe de réponse 64
b. Distorsion harmonique et écrêtage 64
c. Distorsions spectro-temporelles 65
d. Réverbération 65

4. Résistance de l'information sémantique 65

IV.3.2. Squelette sémantique et phonatomes 66

1. Mise en évidence du squelette sémantique 66
2. Squelette sémantique et phonatomes 66
3. Notion de netteté 67
4. Phonèmes et phonatomes 67
5. Anamorphoses du squelette sémantique 68

a. Anamorphose fréquentielle régulière 68
b. Anamorphose fréquentielle irrégulière 68
c. Anamorphose temporelle régulière 69
d. Anamorphose temporelle irrégulière, ou anamorphose rythmique 70

6. Champs de liberté de la voix humaine 70

IV.4. STATISTIQUE PHONETIQUE ET MOTS ALEATOIRES 71

IV.4.1. Statistique phonétique 71

1. Comptage des digrammes phonétiques 71
2. Contraintes phonétiques 73
3. Utilisation des probabilités de digrammes 73

a. Synthèse 73
b. Reconnaissance 74
c. Tests phonétiques 74

...../

IV.4.2. Mots aléatoires 74

1. Principe 74

2. Comptage des digrammes différenciés selon leur position dans le mot 75

IV.5. CONCLUSION 77

V.- LA SYNTHESE DE LA PAROLE 78

V.1. LES OUTILS DE SYNTHESE 78

V.1.1. Machines parlantes et modèles phonatoires 78

1. La machine de KEMPELEN 78

2. La machine de FABER 80

3. Le VODER 81

4. Les analogues électriques du conduit vocal 81

V.1.2. Les synthétiseurs de parole 81

1. Le VOCODER 81

2. Les synthétiseurs paramétriques 83

3. Les relecteurs de sonagrammes 83

a. Le play-back de COOPER 83

b. L'Icophone II à commande optique 84

V.2. LA SYNTHESE PAR PHONATOMES 84

V.2.1. Etablissement d'un premier dictionnaire sur l'Icophone à commande optique 84

1. Nombre de phonatomes utiles 85

2. Schématisation des phonatomes 85

3. Comparaison des dictionnaires 85

V.2.2. Utilisation de l'Icophone III à commande numérique 86

1. L'Icophone III à commande numérique 86

a. Traduction phonétique 86

b. Recherche des phonatomes en mémoire 86

c. Régulation temporelle 87

...../

d. Sortie vocale 87
e. Visualisation 87

2. L'amélioration du dictionnaire 87

a. Le voisement 87
b. Amélioration de certains phonatomes 88

3. Les tests d'intelligibilité 88

a. Listes de mots 88
b. Appareillage 88
c. Auditoire 89
d. Résultats 89

V.2.3. L'Icophone IV et l'anamorphose 89

1. Coefficients d'anamorphose et coefficient d'intonation 90
2. Résultats 90

V.2.4. Comparaison d'un phonatome à tous les autres 91

V.3. LES APPLICATIONS DE LA SYNTHESE 92

V.3.1. La réponse vocale automatique 92

1. Les unités de synthèse classiques 92
2. Applications potentielles des unités de réponse vocale 92

a. Synthèse vocale de messages élaborés par un calculateur 92
b. Prothèse pour muets 93

3. L'Icophone comme unité de réponse vocale 93

V.3.2. La transmission de parole à très faible bande 93

V.4. CONCLUSION 94

VI - LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE 95

VI.1. METHODES ET MOYENS DE LA RECONNAISSANCE VOCALE 96

VI.1.1. Théorie de la reconnaissance des formes 96

1. Position du problème 96

..../

2. Organisation d'un système de reconnaissance 97
3. Théorie de la reconnaissance 97

a. Méthodes de corrélation 97
b. Méthodes paramétriques 97

VI.1.2. Les problèmes d'apprentissage 98

1. Principe du perceptron 98
2. Principe de la matrice d'apprentissage 99
3. Limites de l'apprentissage 100

VI.1.3. Aspects spécifiques de la reconnaissance vocale 101

1. Variabilité du message et adaptation homme-machine... 101

a. Variabilité de la parole 101
b. Adaptation de la machine au locuteur 101
c. Dérive de la voix et adaptation du locuteur ... 102

2. Moyens d'analyse 102
3. Les objets de la reconnaissance 103

a. Reconnaissance analytique, ou phonétique 103
b. Reconnaissance globale, ou "par mots" 104

VI.1.4. Evolution récente de la reconnaissance vocale 105

1. Evolution de la reconnaissance analytique 105
2. Evolution de la reconnaissance globale 106
3. Bilan des recherches actuelles 107

VI.2. CODAGE NUMERIQUE ET ANALYSE SPECTRALE DE LA PAROLE 107

VI.2.1. Définition et enregistrement du corpus de parole 108

1. Signaux de référence 108
2. Grands fichiers 108
3. Petits fichiers 109

VI.2.2. Codage analogique-numérique 109

1. Préparation de l'enregistrement analogique 109
2. Résultat du codage 110
3. Décodage sur IBM 360 du CIRCE 110

VI.2.3. Analyse spectrale 110

1. Détermination de la fenêtre temporelle 110
2. Analyse par FFT 111
3. Retournement du spectre 111

VI.2.4. Pondération fréquentielle 112

VI.2.5. Représentation imprimée 112

VI.3. NORMALISATION DYNAMIQUE ET FREQUENTIELLE DE LA PAROLE 113

VI.3.1. Normalisation dynamique 113

VI.3.2. Normalisation fréquentielle 115

1. Détermination du coefficient d'anamorphose optimum .. 115
2. Utilisation de l'anamorphose en reconnaissance 116
3. Une application de l'anamorphose irrégulière 116

VI.4. RECONNAISSANCE ANALYTIQUE 118

VI.4.1. Normalisation temporelle des phonatomes 118

1. Principe de la segmentation 118
2. Mise au point de la segmentation-normalisation 118
3. Résultats 120

VI.4.2. Apprentissage et reconnaissance des phonatomes 121

1. Apprentissage 121
2. Reconnaissance 122
3. Assemblage en symboles phonétiques 123

VI.4.3. Regroupement des phonèmes en mots 124

VI.5. RECONNAISSANCE GLOBALE 125

VI.5.1. Compression de l'information 126

1. Compression fréquentielle 126
2. Compression d'intensité 126
3. Compression temporelle 126

...../

VI.5.2. Mémorisation et reconnaissance 127

1. Conditions d'apprentissage 127
2. Principe de la reconnaissance 127
3. Résultats 128
4. Variantes 129
5. Efficacité pour un locuteur unique 129

VI.6. DEVELOPPEMENTS ET PERSPECTIVES 130

VI.6.1. Développements de la reconnaissance analytique 130

1. Utilisation du dictionnaire de synthèse 130
2. Recherche rapide en mémoire 131
3. Assemblage de la suite de phonatomes 131

VI.6.2. Développements de la reconnaissance globale 131

1. Mise au point de la méthode de correspondance temporelle 131
2. Application à la parole continue 132
3. Utilisation du dictionnaire de synthèse 132

VI.6.3. Les problèmes sémantiques 132

1. Quelques définitions 133
2. Les trois niveaux de compréhension de la parole 133
3. Acquisition et organisation des concepts 135
4. La compréhension automatique de la parole 135

VII - CONCLUSION 137

ANNEXE I - NOTATION PHONETIQUE 139

ANNEXE II - TAUX DE COINCIDENCE ENTRE DEUX CONFIGURATIONS ... 142

1. Forme binaire 142
2. Adjonction d'un seuil 142
3. Forme générale 143

BIBLIOGRAPHIE ..... 144 à 150

TABLE DES MATIERES ..... 1 à 10